Psychological Bulletin

HARRY HELSON, Editor
University of Texas

CONTENTS

Partial Reinforcement: A Selective Review of the Literature since 1950 ..... DONALD J. LEWIS 1

"Land! Land!" ..... GORDON L. WALLS 29

The Effects of Violations of Assumptions Underlying the *t* Test ..... C. ALAN BONEAU 49

The Influence of Situational and Interpersonal Variables in Projective Testing ..... JOSEPH MASLING 65

Note on the Multitrait-Multimethod Matrix ..... LLOYD G. HUMPHREYS 80

Published Bimonthly by the
American Psychological Association

VOL. 57, NO. 1 JANUARY, 1960

Consulting Editors

R. P. Bonar
University of Texas

W. T. Lambert
Johns Hopkins University

J. P. Guilford
University of Southern California

W. H. Holtzman
University of Texas

D. W. MacKinnon
University of California, Berkeley

L. J. Postman
University of California, Berkeley

S. B. Sells
Texas Christian University

W. A. Wilson, Jr.
University of Colorado

The *Psychological Bulletin* contains evaluative reviews of research literature and articles on research methodology in psychology. This JOURNAL does not publish reports of original research or original theoretical articles.

Manuscripts should be sent to the Editor, Harry Helson, Department of Psychology, University of California, Berkeley 4, California.

Preparation of articles for publication. Authors are strongly advised to follow the general directions given in the *Publication Manual of the American Psychological Association* (1957 Revision). Special attention should be given to the section on the preparation of the references (pp. 50–60), since this is a particular source of difficulty in long reviews of research literature. *All copy must be double spaced, including the references.* All manuscripts should be submitted in *duplicate*, one of which should be an original typed copy; author's name should apppear only on title page. Original figures are prepared for publication; duplicate figures may be photographic or pencil-drawn copies. Authors are cautioned to retain a copy of the manuscript to guard against loss in the mail and to check carefully the typing of the final copy.

Reprints. Fifty free reprints are given to contributors of articles and notes.

ARTHUR C. HOFFMAN
Managing Editor

HELEN ORR
Promotion Manager

SARAH WOMACK
Editorial Assistant

Communications—including subscriptions, orders of back issues, and changes of address—should be addressed to the American Psychological Association, 1333 Sixteenth Street N.W., Washington 6, D.C. Address changes must reach the Subscription Office by the 10th of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

Annual subscription: $8.00 (Foreign $8.50). Single copies, $1.50.

PUBLISHED BIMONTHLY BY:

THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.

Menasha, Wisconsin

and 1333 Sixteenth Street N.W., Washington 6, D.C.

Entered as second class matter at Washington, D.C., and Menasha, Wisconsin. Second-class postage paid at Washington, D.C., and at additional mailing offices. Printed in U.S.A.

Copyright, 1960, by The American Psychological Association, Inc.

VOL. 57, NO. 1 JANUARY, 1960

Psychological Bulletin

PARTIAL REINFORCEMENT:

A SELECTIVE REVIEW OF THE LITERATURE SINCE 1950[1]

DONALD J. LEWIS[2]
Louisiana State University

In May, 1950, Jenkins and Stanley published *Partial Reinforcement: A Review and Critique*, which is now a standard reference in the area. Its very excellence has probably inhibited other workers in the area from attempting another review of the literature. The partial reinforcement literature, however, is so large and has increased at such a rapid rate, that it is difficult now to maintain a perspective on the whole area. The present review is offered, with considerable trepidation, as an attempt to point out at least the major results and trends of the post-1950 research.

Even though the list of references is long, this review unfortunately is not exhaustive. Some studies are not included because they were believed by the writer to be insignificant, inconclusive, or badly conceived. Probably other studies, perhaps excellent ones, simply have been overlooked, although the writer has made a conscientious effort to read every study on partial reinforcement that has been published in major psychological and allied journals. Another large class of excluded studies are those which involve a free responding situation. This exclusion is purely arbitrary and does not mean that such studies are considered to be of less significance, nor that different principles necessarily will be required for their explanation. Only discrete trial studies are considered, primarily because these happen to be of major interest to the writer. A review of free responding studies is badly needed. A very few studies before 1950 are considered either because of their parametric design or their importance to theory.

This paper is organized around the major empirical variables that have been investigated in attempts to determine the effects of partial reinforcement on extinction. The aim in the data section has been to determine parametric relations between stimulus and response variables. Some studies whose primary purpose was to test a theoretical orientation do not allow such a determination, but they are considered to be valuable apart from their theory. Therefore, they, too, are cited in the data section. Without their theoretical context these studies may appear somewhat disjointed, but it is sometimes salutary to look at data only in relation to empirical variables. Other studies seem most important primarily for their contribution to

[1] This report was supported by a grant from the National Science Foundation.

[2] The author wishes to express his appreciation to John W. Cotton of Northwestern University and D. W. Tyler of Louisiana State University who read the entire manuscript. Any inadequacies remaining are not to be attributed to them.

theory and they are considered in the theory section. Some, of course, must be considered in both sections. Because this paper is concerned with extinction, acquisition phenomena are treated only incidentally. Perhaps a necessary introduction to this paper is a thorough reading of the Jenkins and Stanley (1950) paper. The issues and problems they discuss are not taken up again unless there is a clear need.

Data

Percentage of reward. Following their review of the pertinent literature, Jenkins and Stanley (1950) arrived at an empirical generalization which stated: "All other things equal, resistance to extinction after partial reinforcement is greater than that after continuous reinforcement when behavior strength is measured in terms of single responses" (p. 222). Nine years and a great deal of research later this generalization still stands, perhaps more firmly than ever. Because this law of partial reinforcement seems so well established, there seems little point here in noting those studies involving only two percentages of reinforcement whose main point has been to demonstrate again the PRE (partial reinforcement effect). We will be primarily concerned in this section with studies that have attempted to determine a parametric law.

Grant and Schipper (1952) used an eyelid conditioning situation with a light CS and an airpuff UCS and counted the percentage of CRs in acquisition and extinction. The percentages of reinforcement used were 0%, 25%, 50%, 75%, and 100%. The results indicated that the percentage of CRs during acquisition was an increasing function of percentage of reinforcement with the greatest response strength for the highest percentage of reward. During extinction there was a rapid decrease in response strength for the 100% group, and less rapid for the others. The greatest resistance to extinction was for the 50% and 75% groups, falling off for both the 100% and 25% groups. The 0% group showed practically no conditioning and therefore no resistance to extinction. There was, then, a "hard core of resistance" to extinction in the region between, say, 40% and 80% reinforcement.

Duplicating in design the eyelid study discussed above, Grant, Hake, and Hornseth (1951) used a verbal conditioning situation. During acquisition, percentage of positive responses were again an increasing function of percentage of reinforcement, with each group emitting positive responses at about the same rate as it received reinforcements. In extinction, however, the 25% group gave the greatest resistance to extinction, yielding the hump of the ∩-shaped function toward the lower percentages.

Lewis and Duncan (1956b) used a "one-arm bandit" slot machine, modified so that payoffs could be controlled. Each payoff was worth 5¢ to the *S*s, and the percentages used were 100%, 75%, 50%, 37.5%, 25%, 12.5%, and 0%. The total number of plays to quitting was found to be an inverse function of the percentage of reward with the 100% *S*s quitting first and the 0% *S*s quitting last. There was no evidence for a ∩-shaped function.

In another experiment, with 0%, 11%, 33%, 67%, and 100% reward, Lewis and Duncan (1957) asked their *S*s to state for each trial of the 9-trial acquisition series their "expectation" of winning or not winning on the next trial. These expectancies were quantified on a scale from 1 to 6, with 1

representing a firm expectancy of not winning and 6 a firm expectancy of winning. The results showed that expectancies were a regular function of percentage of reinforcement both during acquisition and extinction, and that the expectancy of winning dropped off very rapidly during extinction for the 100% group. This was also the group that quit first. In this case there was a slight drop for the 0% group, suggesting a ∩-shaped function.

Using children from approximately five and one-half to six and one-half years of age in a partial reinforcement situation with plastic toys as reward, Lewis (1952) varied four percentages of reward—100%, 50%, 60%, and 0%—in a 10-trial acquisition series. He found no difference in resistance to extinction between the 50% and 60% groups and between the 100% and 0% groups, although the latter two groups quit significantly sooner than did the former, again a ∩-shaped function.

Five studies have been considered, each of which plotted at least four points along a percentage of reinforcement dimension. The wide variety of experimental situations used has undoubtedly helped to obscure a parametric function. Even so, a ∩-shaped function is found in four of the five studies. Grant and Schipper (1952), Grant, Hake, and Hornseth (1951), Lewis (1952), and Lewis and Duncan (1957) present evidence for such a ∩-shaped function with depressions at both the high and low percentages. Because a nonmonotonic function usually means that at least two processes are operating, Grant and Schipper (1952) guessed as to what these two processes might be. The first process, they hypothesized, is a discriminative one. The higher the percentage of reinforcement, the more the acquisition series should "stand out" from the extinction series, and the less PRE should result. A discrimination process thus results in a decreasing function as a result of percentage of reinforcement. The second process is a learning one. With a response starting close to zero response strength, the greater the percentage of reward, for equal numbers of trials below some limit, the greater the response strength. Thus the learning process produces an increasing function, and the discrimination process should produce a trend in the opposite direction. The combination of these two results in a ∩-shaped function.

If Grant and Schipper are correct, the point of inflection of the ∩ would need to vary with the degree of learning. The greater the degree of learning, the more the point of inflection should move toward the low end of the percentage scale. This is because with a greater degree of learning, the learning process should tend to drop out, leaving only the discrimination process in operation. Several percentages of reinforcement and numbers of acquisition trials need to be combined in the same experiment to verify this conjecture.

Pattern. Not only may reinforcement be given in different percentages, but within any percentage less than 100, the pattern may vary. Within 50% reinforcement, for example, the rewards may be given randomly, irregularly but not randomly, or regularly. And a large number of regular patterns are possible, depending on the length of the acquisition series. A systematic exploration of these variables is needed.

Grant, Riopelle, and Hake (1950) set up random (R), single alternation (SA), double alternation (DA), and 100% groups in an eyelid conditioning situation. At the end of acquisi-

tion the 100% and the R groups were about equal and superior to the DA and SA groups. During extinction the R group showed a very rapid decrement in response magnitude and ended below the other groups. This is one of the few studies reporting faster extinction following partial reinforcement. There is no ready explanation for this paradox.

Longnecker, Krauskopf, and Bitterman (1952) performed an experiment comparing SA and R groups, measuring the galvanic skin response (GSR). They found no statistically significant mean difference between R and SA at the end of acquisition, although the R group appears to have had a stronger response. Very interesting was the saw-toothed acquisition curve for SA. Responses were considerably stronger on reinforced trials than nonreinforced trials. During extinction the R group was clearly and significantly more resistant to extinction than the SA group.

Using a combination elevated runway and a jumping platform with rats as *Ss*, Tyler, Wortz, and Bitterman (1953) explored further the differences between SA and R. Again, there were no differences at the end of acquisition, and again the SA group was significantly less resistant to extinction than the R. Also, the saw-toothed acquisition function appeared for SA, developing about the 60th trial. Faster runs were recorded after a nonreinforced trial. Earlier in acquisition just the opposite was true; faster runs occurred after rewarded trials. A definite patterning effect is thus discernable in these two studies, not only in terms of resistance to extinction, but also in terms of the acquisition function.

Hake and Grant (1951) and Hake, Grant, and Hornseth (1951) were concerned with the patterning effects of *blocks* of reinforced and unreinforced trials. They varied factorially both the number of blocks (and therefore the number of transitions from nonreinforced to reinforced trials) and the number of trials in blocks. This procedure resulted in combining variables, since as the number of transitions increased and as the number of nonreinforced responses per block increased, the total number of nonreinforced trials also increased. In one of the studies (Hake & Grant, 1951) an eyelid conditioning situation was used. No significant results were found. This may be because of the combination of variables mentioned above, or it may be that when working with complete blocks of reinforced and nonreinforced trials, a longer acquisition series is necessary. The acquisition series used here was only 1, 3, and 5 transitions. In a second experiment of identical design (Hake, Grant, & Hornseth, 1951) but with a verbal conditioning situation, both the number of transitions and the number of reinforced trials were significant variables. Three transitions resulted in greater resistance to extinction than either 1 or 5, and the fewer the number of nonreinforced trials the greater the resistance to extinction.

Working with a modification of the Humphreys' board, Grosslight, Hall, and Murnin (1953) gave reinforcements in three different patterns. One group (RR) was reinforced on every trial, a second (RU) received reinforcements in a block always followed by a single block of nonreinforced trials, and the third (UR) had all the nonreinforced trials occurring as a block in the middle of a sequence of reinforced trials. After five acquisition sessions of this kind, an extinction session was given, and the number of "plus" responses determined for the three groups. The UR group was more resistant to extinc-

tion than the other two, and the RU group was more resistant to extinction than RR. A group, it seems, whose acquisition series ends with reinforcement is more resistant to extinction than a group whose acquisition series does not end with reinforcement. A similar finding was reported by Ishihara (1954).

Grosslight and Radlow (1955) obtained similar results with a response reversal situation and albino rats. Response reversal was significantly slower in the UR condition than in the RR and RU conditions. For this study, however, there was no difference between RR and RU. A likely explanation of this latter finding lies in the 24-hour interval separating the sessions in the animal study, whereas the interval was of the order of seconds or minutes for the human study. With a short interval between sessions, the RU group also has a reward closely following nonreinforcement, although not as closely as for the UR condition. For a second study (Grosslight & Radlow, 1957), with rats and a habit reversal problem, only one nonreinforced trial a session for three sessions was used in the RU and UR conditions. Again the UR was significantly slower to reverse, and there was no difference between RU and RR. Again there was a 24-hr. interval between sessions. Finally (Lewis & Duncan, 1956b), one study shows that irregular variations of patterns within a single percentage of reinforcement (25%) results in no difference in PRE.

It seems clear, reviewing the studies on patterning, that SA results in less resistance to extinction than does R (Longnecker, Krauskopf, & Bitterman, 1952; Tyler, Wortz, & Bitterman, 1953). It also seems clear that blocks of nonreinforced trials which end a number of times with at least one reinforced trial are superior to continuous reinforcement and to those blocks which end in a nonreinforced trial (Grosslight & Radlow, 1957; Hake, Grant, & Hornseth, 1951). Little more is known about patterning because even the more obvious variables and their combinations have not been studied.

Secondary reinforcement. In the following discussion, S^n will represent a neutral stimulus serving as a discriminative stimulus which is presented before the response. And S^r will represent the stimulus used as a secondary reinforcer.[3] S^r is presented after the response whose strength it is hoped to strengthen or maintain. Primary reinforcement (S^R) is ordinarily necessary to make a stimulus serve as S^n or S^r, and S^R may be presented according to many different schedules. There are thus three variables, the S^n, the S^r, and the S^R, which may be manipulated somewhat independently, and as a result there are a number of different ways in which secondary reinforcement may be combined with partial reinforcement. Because each of these ways may bring about different results, the problem of the relationship of secondary reinforcement to partial reinforcement fractionates into the number of experimental designs by which the three may be combined.

Three principal experimental designs may be used during acquisition to vary the percentage of presentation of S^R and S^n. Design A-1: Primary reinforcement is given only part of the time and S^n is presented on every occasion. Design A-2: Whenever S^R is given, S^n is also presented, but when S^R is not given, neither is S^n. Design A-3: Primary reinforcement is presented on a certain

[3] No position is implied on the issue of whether a stimulus must first serve as an S^D in order to serve as an S^r.

TABLE 1

Experimental Design	Reinforcer	
	S^R	S^s
A-1	partial	continuous
A-2	partial	partial, contingent on S^R
A-3	partial	partial, not contingent on S^R

percentage of trials and so is S^s, but the two percentages are varied independently so that S^R may coincide with S^s and it may not. These designs are represented in Table 1. Because the effects of "partial secondary reinforcement" can only be determined when compared to "continuous secondary reinforcement," only those studies in which such a comparison is made will be cited in this section.

In addition to these general methods of combining primary and secondary reinforcement are two general methods of presenting S^r during extinction, or during the test trials. The S^r may be presented on every trial—Design E-1—or only on part of the trials—Design E-2. The studies available in this area will be classified according to the design used.

There seems to be only one study in which Design A-1 has been used. Peterson (1956) used a runway for acquisition, and a buzzer was presented on every trial just as the *S* entered a delay chamber. (Peterson was also interested in delay of reinforcement.) Tests for the effectiveness of S^r were given in a bar pressing device in which *S* had had no previous experience. Perhaps this procedure does not fit the usual paradigm for extinction, but as a test for the effectiveness of a secondary reinforcer it is certainly superior to ordinary extinction. Peterson found no difference between an S^r associated with 50% reinforcement and one associated with 100% reinforcement. There is also no evidence that any secondary reinforcement effects occurred with either percentage.

Notterman (1951) varied the number of nonreinforced trials in a runway with the number of reinforced trials held constant. He presented a light as S^s on each reinforced trial just as *S* entered the goal box. On nonreinforced trials S^s was not presented. Thus Notterman used Design A-2. Certain of his groups were then extinguished to the same level (although Notterman's argument for this is not as firm as would be desirable) in the absence of S^r. Then S^r was reintroduced for further trials but no primary reward was given. Response strength increased as a direct function of the number of originally interspersed nonreinforced trials. The smaller the percentage of reinforcement, the greater the effectiveness of the S^r. The smallest percentage of reinforcement used was 33%.

No study has been reported using Design A-3. It is, however, reasonable to guess that if Design A-3 is combined with Design E-2, S^r will be most efficacious in maintaining a response or bringing one about anew.

Three designs for combining partial reinforcement with secondary reinforcement have been considered. For Design A-1, little can be said. Peterson (1956) found no clear evidence for secondary reinforcement even in a standard control group. Notterman (1951), with an A-2 design and a nice parametric study, found greater effectiveness for S^r with smaller percentages of reinforcement. Design A-3 is untried.

Successive acquisitions and extinctions. Perkins and Cacioppo (1950) gave one group of animals 16 ac-

quisition trials, only 8 of which were reinforced, followed by 30 extinction trials. Seven days of this kind of training were given. A second group received the same succession of events except that all acquisition trials were reinforced. The results showed that the 50% *Ss* were more resistant to extinction throughout the extinction sequences, and the rate of extinction increased with successive extinctions.

Lauer and Estes (1955) used a jumping stand and gave three series, each of 14 acquisition and 8 extinction trials. For one group the first acquisition series was with 100% reinforcement, the second with 50%, and the third with 100% again. The other group started with 50% and alternated on successive acquisitions. They found that successive extinctions resulted in *less* extinction, both in rate and terminal level. Extinction rates were similar after 50% and 100%, except that there was a greater decrement on the first extinction day following 100% reinforcement.

Lauer and Estes (1955) pointed out that successive acquisitions and extinctions had a great deal in common with typical partial reinforcement in that nonreinforced trials (extinctions) follow reinforced trials (acquisitions). Razran (1955) has also made this point. The main difference between the two is in the length of the sequence of reinforced and nonreinforced trials, which actually makes successive acquisitions and extinctions identical to the patterning of partial reinforcement. Pursuing this comparison of successive acquisitions and extinctions to pattern of reinforcement, Lauer and Carterette (1957) gave one group successive acquisitions only (A-A) and another group received successive acquisitions and extinctions (A-E). Both groups received spaced trials. The A-A group was presumed to be analogous to 100% reinforcement and the A-E group was presumed to be analogous to 50% reinforcement. They found that the mean starting speed of the A-A group on all reacquisition series was *below* that of the A-E group, even though the former had had more reinforcements. This is a very interesting finding Its typical partial reinforcement counterpart should show that the reinforced trial following a nonreinforced trial would be superior to a reinforced trial following another reinforced trial. Amsel (1958) presented data to indicate that this is what happens. Amsel interpreted his findings to indicate that nonreinforcement has a frustration-drive effect which serves to energize the behavior on the next trial. However, it is hard to see how the frustration-drive could persist over the 30-min. intertrial interval used by Lauer and Carterette (1957). It is just this effectiveness of partial reinforcement procedures on widely spaced trials that is most difficult to interpret.

The Perkins and Cacioppo study (1950) showed decreased resistance to extinction with successive extinctions, while both Lauer studies (Lauer & Estes, 1955; Lauer & Carterette, 1957) showed just the reverse. It is interesting to note that other successive acquisition and extinction studies (e.g., Bullock & Smith, 1953) report findings similar to Perkins and Cacioppo. The reason for the difference in results lies perhaps in the shortness of the extinction series in the two Lauer studies, 8 trials in one case and 12 in the other. The other studies in this area used a much longer extinction period.

In any case, the Lauer studies point the way to a considerable consolidation of partial reinforcement

variables, in that successive acquisitions and extinctions can be considered as an aspect of the pattern and percentage of reinforcement. If any intertrial interval phenomena should turn out to be reliable, they should also show up as a function of the interval between the successive acquisition and extinction series.

Number of trials. Only a few studies have been concerned specifically with the relationship between partial reinforcement and the number of acquisition trials. A large number of studies (e.g., Jenkins & Stanley, 1950) however, have shown that acquisition responding is greater with 100% than with partial reinforcement. A study by Weinstock (1958) amplifies this statement so that it holds only for the early acquisition trials. Weinstock gave 109 acquisition trials and, for the later trials, the smaller percentage groups were performing better than the larger percentage groups, although there was no regular order. During acquisition, then, there is an interaction between percentage of reinforcement and number of trials.

Weinstock reports the usual PRE with relatively clear evidence that rate of decrement for the low percentage groups is less than that for the high percentage groups. In addition, the higher percentage groups achieve a different and lower asymptote than the lower percentage groups. Finally, Weinstock notes a significant increase in performance late in extinction for the 100% group. We are faced here with the anomalous situation that, at least for continuously rewarded groups. rewards depress response strength late in acquisition and nonrewards increase response strength late in extinction.

Lewis and Duncan (1956a, 1958a) in two studies combined different numbers of acquisition trials with different percentages of reinforcement. They found no interaction during extinction for the two variables, but the larger number of acquisition trials, in both cases, resulted in quicker extinction. Capaldi (1957, 1958) in two studies reports a very similar finding; the more acquisition trials, the *faster* the extinction. In his second study, Capaldi combined length of acquisition with pattern of reinforcement. Some *S*s received rewards on every other trial, and some received rewards randomly. The alternately rewarded, long acquisition group extinguished faster than the alternately rewarded, short acquisition group, and faster than either the short or long randomly rewarded groups. There was no difference between the irregularly rewarded groups as a function of the length of acquisition.

The Lewis and Duncan (1956a, 1958a) and Capaldi (1958) studies differ on one point. Capaldi found decreased resistance to extinction after a long acquisition series only for regular reinforcement. Lewis and Duncan, however, found no exception for irregular reinforcement. They stated that for a well-learned response the function of acquisition trials, as far as resistance to extinction is concerned, is to establish a stable stimulus pattern; the more acquisition trials the stabler the pattern, and when extinction begins, the stimulus change will be greater. This should be true with regular and irregular reinforcement, although more acquisition trials would probably be necessary with the latter. Capaldi may not have given a sufficient number of acquisition trials to decrease resistance for his irregularly rewarded group. Also, he presents no evidence that his groups were equal at the end of acquisition.

Spontoneous recovery. There is very

little information about spontaneous recovery as affected by any variable, and certainly there is little known about the effect of partial reinforcement on spontaneous recovery. Perkins and Cacioppo (1950), whose study has been briefly described above, comment simply, "Spontaneous recovery was complete before each reconditioning trial." Notterman, Schoenfeld, and Bersh (1952) conditioned heart rate by means of electric shock. The partial group showed no evidence of extinction although the continuous group did. No spontaneous recovery was possible for the partial group, nor was any found for the continuous group.

Lewis (1956) with a straight alley found no differential effect of partial reinforcement on spontaneous recovery; spontaneous recovery occurred for both partial and continuous groups. Lauer and Estes (1955) found spontaneous recovery occurring from day to day following 100% reinforcement, but not following 50% In fact, there was a decrement in response between daily blocks for the 50% group, but recovery seemed to occur within the daily blocks.

The empirical evidence is too meager and conflicting to warrant any firm conclusion about the effects of partial reinforcement on spontaneous recovery.

Punishment. In the avoidance conditioning situation, *S* can avoid getting punishment by an appropriate response. Thus, early in acquisition, *S* gets punished on some of the trials and not on others, a partial procedure. In the escape learning situation, *S* gets the punishment no matter what he does, and this is a continuous procedure. If, as is commonly conceived, shock offset is reinforcing, one would expect, from the principle of partial reinforcement, that extinction would be more prolonged after avoidance conditioning than after escape conditioning. Jones (1953), Logan (1951), and Sheffield and Temmer (1950) have shown this to be the case. Jones (1953) has also shown that an "intermittent escape" schedule, one that involves withholding punishment on some trials, results in greater resistance to extinction than orthodox escape. Wynne and Solomon (1955) have pointed out, in addition, that the avoidance *extinction* situation actually involves further learning of the instrumental response. They argue that two responses are learned during the acquisition series, one is the instrumental response and the other is a conditioned emotional response (CER). The CER has a longer latency than the instrumental response, and on most trials the instrumental response occurs before the CER can begin. The *S* is removed from the conditioned stimulus before the CER can be evoked. On some trials, however, the latency of the instrumental response is sufficiently long for the CER to occur, and thus the instrumental response serves to reduce the CER, and the instrumental response is reinforced. Because this type of reinforcement occurs on only some of the trials, the instrumental response is actually partially reinforced during extinction.

Intertrial interval. Sheffield (1949) reported evidence to indicate that the PRE could be obtained if the acquisition interval was massed (15 sec.) but not if it was spaced (15 min.), and gave an "aftereffects" interpretation of her results. Since her conclusion is based on a presumed interaction between percentage of reinforcement and acquisition interval, the analysis of variance was the appropriate statistical technique. Her analysis by means of *t* was, how-

ever, suggestive of her conclusion.

Two attempts to replicate the Sheffield study have been reported by Wilson, Weiss, and Amsel (1955). They tried to manipulate the strength of the aftereffects of rewarded trials by using dry food in one experiment and water in the other; Sheffield had used wet mash. The results obtained by Wilson, Weiss, and Amsel were in sharp disagreement with those of Sheffield. The 50% reinforcement conditions led to greater resistance to extinction independent of the acquisition interval. There was no significant interval by reinforcement interaction either by analysis of variance or analysis of covariance. Lewis (1956) also replicated Sheffield's experiment except that he used a 2 min. distributed interval. He obtained essentially the same results as those of Wilson, Weiss and Amsel. Again the PRE was obtained whether or not the acquisition trials were distributed or massed.

With an eyelid conditioning apparatus, Grant, Schipper, and Ross (1952) used a 2×2×2 factorial design similar to that of Sheffield except that the intertrial intervals were 10 sec. or 40 sec. The extinction results were quite complex in that there was a triple interaction among the three variables of the experiment. The authors interpreted this interaction to mean that the superiority of the 50% groups during extinction was increased when there was a change in the distribution of trials from acquisition to extinction.

The Sheffield design was also used by Grant, Hornseth, and Hake (1950) in a light-guessing experiment using a Humphreys' board (Humphreys, 1939). The intertrial intervals were 5 sec. and 45 sec. Again the PRE was obtained independent of the acquisition interval. Tyler (1956) in a discrimination situation found the PRE after a 15 min. acquisition interval, and Weinstock (1954, 1958) in two studies found the PRE after a 24-hr. interval between acquisition trials. The only study not showing the PRE after spaced acquisition trials was that of Rubin (1953), but it is impossible to separate the effects of the intervals from possible other effects such as those of secondary reinforcement, and little can be concluded from this study.

The conclusion seems to be rather firm that the PRE obtains whatever the spacing of the acquisition trials.

Drive. There are only two papers directly relevant to the effects of different drive levels on partial reinforcement. One by Lewis and Cotton (1957) is essentially negative. Using 1, 6, and 22 hr. of food deprivation with 50 and 100% reinforcement for each drive level, they found that the interaction between the two variables did not attain significance.

The other study relevant to this variable is that of Linton and Miller (1951). Two groups of hungry rats received acquisition training with either 100% or 50% reinforcement. Each of these two groups was subdivided into two groups during extinction, one of each pair being extinguished while satiated, the other under the acquisition drive. Although appropriate for a factorial analysis of variance, the data were analyzed by *t* tests. The results indicated that the group partially reinforced and extinguished under the same drive as used during acquisition was more resistant to extinction than the other three, which were not different from each other. In part, the results were interpreted in terms of drive-stimulus generalization. Another part of the explanation had to do with the absence of a frustration

drive for the partial-satiation group. It was believed that the absence of the frustration drive for the satiation group prevented the PRE from appearing.

Because a motivational interpretation of the PRE is not uncommon (Amsel, 1958) more studies are needed combining drive with different schedules of reinforcement.

Discrimination. In one sense, all studies using a choice situation and the noncorrection method involve partial reinforcement, because an incorrect trial is never reinforced. As the number of correct trials increases, so does the percentage of reinforcement. Lewis and Cotton (1958) have pointed out the complexity of the T maze as far as partial reinforcement is concerned. The two arms and the stem of the T maze each have a *different* percentage of reward obtaining. The correct arm is rewarded 100% of the time. The incorrect arm is rewarded 0%, and stem is rewarded at varying percentages, low during the early trials and almost 100% as the response is learned. Discrimination situations have thus afforded a difficult arena for the study of partial reinforcement. But because choice situations are so commonly used and because partial reinforcement is such an effective variable, studies are needed to tease out the effect of the different percentages of reward in the different sections of the apparatus. At a minimum, extinction running speeds should be determined separately for the stem and two arms of the T maze. It might also be desirable to duplicate in a runway the pattern of reinforcement usually obtained in the stem of a T.

Some information about partial reinforcement is available from response reversal studies, even though response reversal does not qualify as an orthodox extinction operation.

Wike (1953) showed that a response acquired under conditions of partial reinforcement is more resistant to reversal training than a continuously reinforced response. This finding has been confirmed by Grosslight and Radlow (1955), Grosslight, Hall, and Scott (1954), and Kendler and Lockman (1958), but not by Buss (1952).

Babb (1956) has shown that an irrelevant stimulus associated with reward 70% of the time facilitates later performance when the irrelevant stimulus becomes relevant. When the irrelevant stimulus is associated with reward 50% of the time, it has no effect on later learning. When associated with reward 30% of the time, later learning is inhibited by this previously irrelevant stimulus.

In general, partial reinforcement retards later response reversal.

Generalization and complex stimuli. In a study investigating the gradient of primary stimulus generalization to tones, Wickens, Schroder, and Snide (1954) established conditioned galvanic reactions to one of three tones, each of which was separated by 25 jnd's. For one group, in addition to the tones, a click was randomly presented 12 times without reinforcement.

During test trials a gradient of generalization was found for the group without clicks, but no gradient was discovered in the group with clicks. The failure to obtain a gradient for the click group seemed attributable to a generally high resistance to extinction.

Other evidence is available (Brown 1947) to indicate that the generalization gradient is very broad on the first test trials, becoming sharper with further test trials. Because the test trials are usually nonreinforced,

generalization, and of course discrimination, is at least a by-product of extinction. A variable which increases resistance to extinction should, then, decrease any generalization gradient. From this, one would conclude that partial reinforcement should increase the amount of generalization, which is indicated by the Wickens, Schroder, and Snide (1954) study. Further evidence on this important problem is needed from other experimental situations.

The experimenters later reasoned that "if a complex stimulus is presented along with reinforcement and an element of the complex is presented at some other time without reinforcement, then this latter action increases the resistance to extinction of the complex stimulus itself" (Wickens & Snide, 1955, p. 257).

An experiment (Wickens & Snide, 1955) was undertaken to test this hypothesis directly. The general procedure consisted of employing as the CS a light and a tone which came on simultaneously. Two groups were conditioned and extinguished to this stimulus complex. For the experimental group one or the other aspect (the light alone or tone alone) of the total complex was occasionally presented during the training without reinforcement. According to the hypothesis, the experimental group should show greater resistance to extinction than the control. This hypothesis was supported by the data.

Magnitude of reward. Lewis and Duncan (1956b) combined different magnitudes of reward with partial reinforcement using a modified slot machine. Their rewards were 1, 10, 25, and 50 cents for each rewarded acquisition play. There was no significant interaction between amount of reward and the five different percentages of reward used in this situation. Hulse (1958), however, with an enclosed alley and magnitudes of .08 gm and 1.0 gm pellets, found a larger PRE with the larger sized pellet. The two experiments differ so greatly that it is almost futile to guess at the crucial difference. One possibility, however, lies in the intertrial interval which Hulse believes important for reward-percentage interactions. Hulse used a 24-hr. interval and Lewis and Duncan used a massed trial situation. Also, there may again be a difference between a *performance* situation like Lewis and Duncan's and essentially a *learning* one like Hulse's.

Confinement. In addition to magnitude Hulse also varied the length of the goal box confinement. He found that the PRE was less if there was a *change* in confinement from acquisition to extinction. Different extinction confinements had no effect, confirming a previous study by Hulse and Stanley (1956). Acquisition confinements were not effective by themselves either.

Variable delay of reward. Crum, Brown, and Bitterman (1951), in testing one of the theoretical notions of Sheffield (1949), introduced a new and interesting variable into the partial reinforcement literature. For one group of *S*s, reward was given immediately. For a second group, reward was given immediately on half of the trials and it was delayed for 30 sec. on the other half. The trials on which reward was delayed were irregularly interspersed among the others. They found the typical PRE for the variable delayed group. Confirmation was soon made available by Scott and Wike (1956). Logan, Beier, and Kincaid (1956) showed that the PRE occurred when delays were variably 0 and 30 sec.,

but not when they were 0 and 9 sec. Kintsch and Wike (1957) used a **T** maze and gave one group immediate reward on the correct trials, another group 10-sec. delays on half of the correct trials, and a third group 30 sec. delays on half of the correct trials. The results were essentially the same as those of Logan, Beier, and Kincaid (1956); and PRE appeared with 30 sec. variable delay, but not with 10 seconds. Feher (1956), however, had found increased resistance to extinction with constant acquisition delays of 10 seconds.

Wike and McNamara (1957) used an alley runway, and varied percentage of delay. One group was delayed on 25% of the trials, a second group was delayed on 50% of the trials, and a third group was delayed on 75% of the trials. All delays were 30 seconds. On the other trials *Ss* received reward immediately. The 25% group extinguished faster than the other two, which were not significantly different. Apparently, reward must be delayed on more than 25% of the trials for the PRE to occur. It would be interesting to know the shape of the function between 25% and 50%.

Peterson (1956) combined partial delay with partial reinforcement in an interesting experiment. One group was reinforced on every trial with no delay. Another also had 100% reinforcement with delays on various trials of 0, 10, 20, and 30 sec. A third group had 50% reinforcement with delays of 0, 10, 20, and 30 sec. on the reinforced trials. A final group had 50% reinforcement with no delay on the reinforced trials. The 50% group with delay was most resistant to extinction. The 100% group with delay and the 50% group without delay were about equal and next most resistant to extinction. Least resistant to extinction was the 100% no delay group.

In conclusion, it seems that partial delays of 30 sec. result in increased resistance to extinction, but partial delays of 10 sec. or less do not. The delays must occur on more than 25% of the trials, and when partial reinforcement is combined with partial delay, resistance to extinction is the greatest.

Situations. Having established an empirical phenomenon in one experimental situation we take some practical, and sometimes theoretical, interest in determining th variety of situations in which the phenomenon holds. In this section an attempt will be made to review quickly some of the different experimental situations in which the PRE has been explored.

Lewis (1952) used children in a button pushing situation. The rewards were plastic cowboys and, in a 10-trial acquisition series, were given 100%, 60%, 50%, and 0% of the time. There was no difference in PRE between the 50% and 60% groups, nor between the 100% and 0% groups, but the former two groups were superior to the latter two. Fattu, Mech, and Auble (1955) with a similar situation measured the number of responses to extinction after 100%, 50%, and 25% reinforcement. The 25% group was considerably the most resistant to extinction, followed by the 50% and 100% groups in that order. Fattu, Auble, and Mech (1955) were not able to repeat their results although the same trend was apparent.

Goss and Rabaiola (1952) had *Ss* learn one of three nonsense syllables to a color stimulus. The correct response was reinforced with a buzzer 100%, 75%, and 50% of the time for the different groups. The speed of attainment of a criterion of learn-

ing was an increasing function of the percentage of reward. After the criterion was reached the reward conditions were switched. Half of each acquisition group was given 0% reward, and the reward percentage was halved for the other. The mean number of correct responses decreased faster for the 50% and 0% group than for the 100% to 0% group. This runs counter to the usual PRE. Also, Goss and Rabaiola report an increase in correct responses after about eight trials of no reward.

Hirsch (1957) presented his *Ss* with a series of words to which they were to respond with a number. The correct number was one less than the number of letters in the words. He reinforced his *Ss* either 100% or 67% of the time for the correct responses. All reinforcement was omitted after a criterion was attained. Again there was no significant PRE. Perhaps, if more of a "persistence" measure of extinction had been used in these two studies, the typical PRE would have occurred.

Kanfer (1954) and Spivok and Papajohn (1957) varied percentage of reinforcement in an autokinetic situation. A restricted class of *S*'s verbal statements about movement was selected for reinforcement. Kanfer rewarded these responses at 100% 67%, 50%, and 0% for different groups, bringing all but the 0% group to the same criterion of performance. The reward was removed. The typical PRE was found in that the greater number of "critical" responses was found in the partial group. The *S*'s awareness of the reinforcement was not an essential condition for the PRE. Kanfer also pointed out that the rate of extinction may be as important as the total number of responses emitted. Spivok and Papajohn (1957) compared 100% reinforcement with a variable interval schedule, finding greater PRE for the latter.

Finally, Lewis and Cotton (1958) found the PRE in a T maze with a correct turn response measure. Their animals had partial reinforcement administered by being placed in the goal by hand with food present on 50% of the placements. However, both the 50% and 100% groups were worse than a control group that did not receive nonresponse reinforcement.

Theory

In 1950, discussing the various theories of partial reinforcement, Jenkins and Stanley (1950) enumerated five essentially different theories. These were: (*a*) response unit, (*b*) aftereffects, (*c*) discrimination, (*d*) secondary reinforcement, and (*e*) expectancy. All of these remain today as viable entries in the partial reinforcement sweepstakes although the response-unit hypothesis is rarely used for discrete trial phenomena, and the expectancy hypothesis has been much battered of late and seems to have no articulate supporter remaining. In addition, there are two new entries since 1950. These are: (*a*) a competing response theory of Weinstock (1954), and one by Hulse and Stanley (1956), and (*b*) a mediating response theory, espoused by Amsel (1958), Kendler, Pliskoff, and D'Amato (1957) and Logan, Beier, and Kincaid (1956) among others.

Aftereffects. One of the first, and probably still as good as any, statements of the aftereffects theory is that of Sheffield (1949). She pointed out that the aftereffects of reinforcement are quite different from the aftereffects of nonreinforcement. When reinforcement occurs, *S* would have a

food taste in the mouth, perhaps food particles, and still other stimuli associated with eating. After a nonreinforced trial the aftereffects would include frustration, searching, etc. Obviously the stimuli following nonreinforcement would be very different from those following reinforcement. If the stimulus aftereffects of nonreinforcement are still present on the next trial, and the next trial results in reinforcement, then the instrumental running response would be conditioned to the aftereffects of nonreinforcement, and the *S* would actually learn to respond to the stimuli of nonreinforcement. Since the extinction stimuli are those of nonreinforcement, *S* has learned to respond during extinction. Those *S*s who are rewarded on every trial never have the opportunity to respond to nonreinforcement cues and thus have not learned to respond to extinction stimuli.

But this is only the learning factor of the aftereffects theory as Sheffield describes it. There is also a primary generalization factor. For the continuously reinforced *S*s, the advent of extinction introduces new stimuli for the first time, those from nonreinforcement. With the introduction of new stimuli there will be a response decrement due to primary stimulus generalization. For the partially reinforced *S*s there would be no new stimuli introduced with extinction, since nonreinforcement had occurred repeatedly during the acquisition.

To demonstrate the adequacy of her reasoning, Sheffield ran 50% and 100% groups, factorially combined with massed and distributed acquisition trials. With massed trials the aftereffects of nonreinforced trials should still be present at the initiation of the succeeding trial, as the theory requires. With distributed trials this would not be the case. Her results probably confirmed her theory for the PRE seemed to appear, as tested by *t*, only after massed trials and 50% reinforcement.

Following some implications of Sheffield's hypothesis, Grosslight and Radlow (1955) and Grosslight, Hall, and Murnin (1953) found that several series of trials in which a single nonreinforcement is followed by reinforcement would result in the PRE, if the nonreinforcement were followed by a reinforcement. The authors interpret their results in terms of a Sheffieldian aftereffects theory, in that the aftereffects of the nonreinforced trials are presumed to be conditioned to the instrumental response because they are followed by a reinforcement, but it should be noted that a discrimination hypothesis would handle the results equally well.

Linton and Miller (1951) found no PRE when extinction, after partial reinforcement under normal deprivation conditions, was carried out under conditions of drive satiation and with the reward present on each trial. They reasoned that this was because there were very different aftereffects occurring during satiated "extinction" and that the instrumental response had not been conditioned during acquisition to these aftereffects.

As noted in the section on intertrial interval, Wilson, Weiss, and Amsel (1955) repeated Sheffield's study twice, with minor variations and without her results. In one case the variation was such as to enhance the aftereffects. In the other, the variation served to reduce the aftereffects, but in neither case was the massed acquisition-partially reinforced group more resistant to extinction. Lewis (1956) also repeated

Sheffield's study using a 2-min., instead of a 15-min. spaced interval. The graph of his results appeared to be Sheffieldian, but the appropriate statistical analysis showed no significant effect.

Continuing the attack on the aftereffects hypotheses Tyler (1956) found the PRE in a discrimination situation even after a 15-min. acquisition interval and, even more devastating, Weinstock, on two occasions (1954, 1958) has found the PRE even when acquisition trials were spaced 24 hours apart. This suggests that long range aftereffects are more important than previously believed.

Tyler, Wortz, and Bitterman (1953) reasoned that if Sheffield were right, a simple pattern—alternating reinforcement with nonreinforcement—should give a greater PRE than a random pattern, because alternation maximizes the number of times nonreinforcement follows reinforcement. Their results showed just the opposite; alternation resulted in quicker extinction than did the random pattern.

Another ingenious attack on the aftereffects theory was performed by Crum, Brown, and Bitterman (1951). If the reward were delayed after each trial, they argued, the aftereffects for the succeeding trial would always be those of reinforcement. They compared a group which received delay of reward on some of its trials and immediate reward on others with a group that received immediate reward on all trials. The partial delay group showed the PRE very nicely. Peterson (1956), Logan, Beier, and Kincaid (1956), and Wike (1953), have confirmed the results of Crum, Brown, and Bitterman.

Feher (1956) performed some interesting variations on the delay of reward. She compared delays given before the reward with delays given after the reward. For the latter, *Ss* were simply left in the goal box for a specified period of time. She found that both delay groups extinguished slower than groups without delay. And the delay-before-reward group extinguished slower than the delay-after-reward group, again counter to an aftereffects notion.

Katz (1957) ran three groups of animals in a test of the aftereffects notion. The acquisition trials were given in two quite different runways. One trial was given on the first runway, followed immediately by a trial on the second runway. One group (P-C) was given partial reinforcement on the first runway and continuous reinforcement on the second. A second group (C–P) was given just the reverse: continuous reinforcement on the first and partial on the second. The final group (C-C) was given continuous reinforcement on both runways. For the P-C group, all nonreinforced trials occurring in the first runway were followed by reinforced trials in the second runway. For the C-P group there was no similar opportunity for conditioning the aftereffects of nonreinforcement. All extinction trials were given in the second runway. The results showed that the C-P was slowest to extinguish, followed by the P-C group, and the C-C group extinguished most rapidly. This would indicate that other factors than aftereffects are most important in bringing about the PRE.

The only conclusion that can be drawn from these experiments is that, at best, Sheffieldian aftereffects are not important contributors to the PRE, and they probably have no effect whatsoever.

Discrimination. The discrimination theory, which was apparently first advanced by Mowrer and Jones

(1945), has had the advantage of being less specifically stated than the aftereffects theory, and the longevity of a theory is apparently inversely related to the specificity with which it can be stated and tested. In general the discrimination theory states that resistance to extinction is a function of the similarity of the acquisition stimuli to the extinction stimuli. The more similar the stimulus conditions are in the two situations, the greater the resistance to extinction. The problem then becomes one of stating and demonstrating the variables of which similarity may be said to be a function. Once lawful relationships are obtained between similarity variables and behavior, any controversy over whether discrimination theory is a perceptual one, or refers to something going on in the rat's mind, is superfluous. The major task for discrimination theorists, then, is one of stating specifically what variables determine similarity.

Probably the most vigorous supporters of the discrimination hypothesis have been the Texas group, although their research has been aimed more at disproving the aftereffects theory than it has at giving support to a discrimination theory. Longenecker, Krauskopf, and Bitterman (1952) and Tyler, Wortz, and Bitterman (1953) showed that a simple alternating pattern of reinforcement and nonreinforcement resulted in quicker extinction than a random pattern. Since this was contrary to an aftereffects theory, they argued, there must be come serial patterning that occurs, enabling *Ss* to discriminate the acquisition series and to stop responding quickly when it ceases.

Bitterman, Fedderson, and Tyler (1953) and Elam, Tyler, and Bitterman (1954) gave rewards in one end box and nonrewards in a very different one. One group was extinguished with the rewarded end box present and another group was extinguished with the nonrewarded box. Presumably, the stimulus situations from acquisition to extinction would be more similar when the nonrewarded end box was present because these stimuli had always in the past accompanied nonrewards. Using the rewarded end box during extinction would bring about a stimulus change because these stimuli had always accompanied rewards. The predicted results occurred, and were interpreted according to the discrimination theory. The authors tended to speak as if the discrimination were perceptual in some fashion, but, of course, all they could do was to relate behavior to stimuli, in this case stimulus change. There is no need to introduce an intervening perceptual process to handle this data.

Monkeys were given their choice, by Elam and Tyler (1958), between two stimuli, A and B. For one group, the A stimulus was rewarded 60% of the time and the B stimulus was rewarded 40% of the time. For the second group, the A stimulus was also rewarded 60% of the time but the B stimulus was rewarded 0% of the time. During extinction, the A stimulus was rewarded 0% of the time and the B stimulus was rewarded 100% of the time for both groups. Results showed that during acquisition, A was much more preferred by the second group, but during extinction, A extinguished much more rapidly for the second group. Elam and Tyler attribute the much faster extinction of A for the second group to the greater stimulus change from acquisition to extinction occurring for this group than the other.

Goodnow and Pettigrew (1956)

presented a somewhat similar problem to four groups of human *S*s. The apparatus was a "two-armed bandit" One group received reward for a LR (left-right) pattern of responses, the second received reward for an LL pattern, a third was given an LLR pattern, and the fourth was given a random pattern. After an acquisition series of trials, all groups were given a random pattern for a while and then all groups were switched to a final LR pattern. Both the initial LR and LL groups learned the final pattern faster than the random group. (Many of the LLR group did not learn the initial pattern.) Goodnow and Pettigrew concluded that the final pattern was made more discriminable because of the stability of the initial response pattern. Also, there was less of a stimulus change for the two initially systematic groups than for the initially random group.

McClelland and McGown (1953) gave one group of animals irregular reinforcement in various parts of a circular shaped apparatus. Another group always found the reinforcement in the same place. Results showed that the inconsistently rewarded group was more resistant to extinction. The authors concluded: "Omission of reinforcement changes the cue pattern less if the original cue pattern is complex or variable, than if it is simple and invariant."

Sheffield and Temmer (1950) have shown that the superior resistance to extinction of avoidance learning to escape learning can be given a partial reinforcement interpretation. Jones (1953) found that an "intermittent escape" group, one that was not punished on every trial but still could not learn to avoid, was more resistant to extinction than an orthodox escape group. Also a "limited avoidance" group extinguished slower than the escape group. "Limited avoidance" referred to a condition in which the *S* was placed on an initially uncharged grid, but the current would be turned on before *S* could get off the grid. The faster *S* moved the less shock he would take, and in this sense it was an avoidance condition, but even so, *S* would always get the punishment. The "limited avoidance" group was also more resistant to extinction than the pure escape group. The acquisition conditions, according to Jones, were more like the extinction conditions for the "intermittent escape" and the "limited avoidance" conditions than they were for the escape conditions.

In several studies Lewis and Duncan (1956a, 1956b, 1957, 1958a, 1958b) found that 0% reinforcement showed more resistance to extinction than 100% reinforcement, and more than most other percentages of reinforcement. Their situation, a "one-armed bandit," differed from most others in that no learning was involved in the performance of the lever pulling response. All *S*s, on coming into the experimental room, already knew how to pull a lever. The experimenters argued, therefore, that the acquisition series served primarily to set the extinction series apart, and that 0% acquisition was most like the extinction series. In one of these studies (1958a) they found, contrary to most previous results, that extinction was quicker after a long acquisition series. Again they attributed this to the pure performance situation in which a relatively long acquisition series would serve to make the initial stimulus situation more stable.

Somewhat similar results have been reported by Capaldi (1957, 1958).

Brand, Woods, and Sakoda (1957) in a two alternative situation varied both the difference in percentage of reinforcement and the difference in the ratio of reinforcement. The difference, for example, between 57% and 25% is 50%, and the ratio is three to one. The difference between 100% and 50% is also 50%, but now the ratio is two to one. They found that extinction is quicker when the percentage difference is great (a finding in a line with a discrimination hypothesis), but that the ratios had practically no effect.

Fehrer (1956), who found that delay of reinforcement leads to increased resistance to extinction, also interprets her study according to a discrimination theory in that having a delay come before a reward affords an experience—that of no reward—much like the initial experience during extinction.

All of the studies that have been reviewed in this section so far present evidence that has been interpreted as favorable to discrimination theory. There are two recent studies, however, that present a considerable hurdle for discrimination theory to jump. Marx (1958) gave two groups of animals acquisition training in a runway with food pellets presented in the end box in a glass cup. During extinction one group was presented with the empty cup on every trial, and one group was presented with the empty cup on only half of the trials; on the other half the food cup was absent. There was a greater stimulus change for the group that had the cup on only half of the trials, yet this group extinguished slowest. In a more elaborate experiment, Brown and Bass (1958) varied the stimulus conditions—different alleys—in both acquisition and extinction. They found that variable extinction conditions resulted in increased resistance to extinction. In both of these experiments, stimulus change increased resistance to extinction instead of decreasing it as the discrimination hypothesis would seem to demand, but, it should be noted, only 100% reinforcement was used in both studies.

In addition, Brown and Bass found that irregular stimulus conditions during acquisition had no effect on resistance to extinction. This seems to run counter to previous evidence (MacKintosh, 1955), but Brown and Bass have the better designed study in that theirs was a factorial combination of variables in both acquisition and extinction. MacKintosh extinguished all *S*s under constant conditions, making it impossible, as Brown and Bass point out, to distinguish between learning condition irregularities and the change from irregular acquisition to constant extinction. Both the Brown and Bass and the Marx study are consistent with much previous evidence (e.g., Glazer, 1958) indicating that stimulus novelty and change increases performance. It seems, then, that one must at least conclude that stimulus change during extinction must be weighed more heavily in increasing resistance to extinction than constancy of stimulus conditions from acquisition to extinction. In any case, the studies of Brown and Bass, (1958) and Marx (1958) complicate the task of discrimination theorists.

Secondary reinforcement. Denny (1946) is usually given credit for the formal introduction of the secondary reinforcement hypothesis into the area of partial reinforcement. He pointed out that on reinforcement trials the stimuli of the goal box are associated with primary reinforcement and therefore should acquire secondary reinforcing power. On

nonreinforced trials and during extinction, secondary reinforcement should be taking place which would retard extinction.

The attack on the secondary reinforcement hypothesis began with Bitterman, Feddersen, and Tyler (1953) and Elam, Tyler, and Bitterman (1954). They found that rats extinguished in the goal box in which they had received their reinforced trials—the nonreinforced trials were given in a very different goal box—showed less PRE than those extinguished in the goal box which was present for nonreinforced trials. According to a Hullian (1943) secondary reinforcement hypothesis, the results should have been just the opposite. The stimuli associated with primary reward and present during extinction should have prolonged extinction performance. In only slightly different situations Notterman (1951) and Freides (1957) have duplicated the significant aspects of these studies. They found no difference between the two relevant groups, but this still does not support a secondary reinforcement interpretation.

Mason (1957) has reported an interesting experiment. His animals learned two discriminations. In one, the positive stimulus was rewarded 100% of the time; in the other it was rewarded 50% of the time. On the test trials both positive stimuli were presented and the *S*'s preference on successive trials constituted the response measure. Most *S*s chose the stimulus that had been rewarded 100% of the time, even when twice as many total trials had been given to the 50% rewarded stimuli, so that the total number of primary rewards in the two situations was the same, but the 50% group would have had secondary reinforcement in addition. These results were interpreted as opposed to a secondary reinforcement hypothesis.

The results of Fehrer (1956) can also be interpreted as opposed to a secondary reinforcement interpretation of partial reinforcement. Two of her groups involved goal box delays. One was given a 30-sec. delay before a 10-sec. eating period, and one was given a 10-sec. eating period followed by a 30-sec. delay. Another group was allowed to eat for a full 40-sec. in the goal box. The goal box stimili had a longer association with primary reward for the latter group and should have resulted in slower extinction than the other two. Just the opposite happened; both delay groups extinguished slower than the nondelay group.

Hulse and Stanley (1956) reported, but only at the .10 level of significance, that the PRE occurred only when the S^r was present on every training trial.

The evidence, at this point, seems quite conclusive that secondary reinforcement is not the sole explanation of the PRE, but that it can be exceptionally efficacious was indicated by Zimmerman (1957), who used a free responding situation. He suggested that when an S^a is irregularly associated with reinforcement, it loses none of its S^a function and maintains its S^r function over a considerable time. It is most effective as an S^r when presented irregularly again. Zimmerman's study is based on very few animals, and needs to be replicated before it can be accepted with absolute confiidence, but it seems probable that a stimulus, when irregularly associated with reinforcement and then used irregularly as an S^r, can have very powerful reinforcing value. At this point, however, it becomes difficult to distinguish between a discrimination hypothesis

and a secondary reinforcement one. It seems likely, in fact, that they are basically the same.

Competing response. Weinstock (1954) has presented an "habituation" theory which states essentially that nonreinforced responses "habituate" by some unspecified process and drop out. For a group of partially rewarded *Ss*, competing responses are made during acquisition on the nonreinforced trials, and these competing responses drop out. Thus when extinction begins, the competing responses have already habituated and the instrumental response continues strongly. No habituation of the competing responses occurs during acquisition for the continuously reinforced group, and their onset during extinction results in a rapid decrease in the instrumental response.

The only experiments aimed directly at testing this notion—although Tyler (1956) has pointed out that Weinstock's theory does not explain why random reinforcement results in a greater PRE than alternating reinforcement—are those of Stanley and Clayton (1955) and Hulse (1958). Stanley and Clayton assumed that if *Ss* were delayed in the goal box, more competing responses would occur and thus habituate. They gave immediate reinforcement and a 30-sec. delay to two groups during acquisition and broke these two down factorially to immediate goal box removal and a 30-sec. delay during extinction. The acquisition delay group was not more resistant to extinction than the immediately rewarded group. These results are opposed to Weinstock's habituation theory, if habituation is a function of time in the goal box, but they are also opposed to the results of Fehrer (1956) who found that 30-sec. delays, both before and after the reinforcement, retarded extinction.

In another study, Hulse and Stanley (1956) presented a theory somewhat similar to that of Weinstock (1954). They also argued that competing responses occur during the nonreinforced acquisition trials, but for them the competing responses do not "habituate." The *Ss* learn to do something other than eat during these trials. When extinction starts, the partially reinforced *Ss* do the "something else" that they learned during acquisition and are quickly removed from the stimuli which evoke eating, and thus the conditioned eating response is protected from rapid extinction. At this point Hulse and Stanley follow the notion of Sheffield, Roby, and Campbell (1954), that the faster extinction of conditioned eating leads to a faster loss of the instrumental response. This is because the conditioned eating response works backward over the maze stimuli and constitutes an important part of the stimulus situation leading to the instrumental response.

Hulse and Stanley's theory appears to be more specific than Weinstock's because it does not contain such a vague term as "habituation" playing an important role. Also Hulse and Stanley indicate how the goal box competing response can affect behavior occurring at the beginning of the apparatus. But this notion, too, runs into difficulty when Freides' (1957) data are considered. He found that behavior in the goal box (approaching food) could extinguish while a runway response remained strong.

Mediational responses. It should be noted that the theory of Hulse and Stanley (1956) involved a mediating

conditioned eating response and could very properly be considered in this section also.

Wilson, Weiss, and Amsel (1955) and Amsel (1958) have argued for a mediating frustration response as an explanation of the PRE. During partial reinforcement an emotional response develops on the unreinforced trials. The emotional response works backward over the maze stimuli and is evoked by the stimuli of the start box. The emotional response has stimulus properties, as do all responses, and these stimuli become conditioned to the instrumental response. Since the emotional response does not get conditioned during acquisition to the instrumental response for the 100% *S*s, they show a sharp response decrement during extinction.

Essentially the same idea is presented by Kendler, Pliskoff, D'Amato, and Katz (1957), except that no special properties, such as emotionality, are ascribed to the mediating response. Lewis and Duncan (1958b) also used a mediating response to interpret their data and show how some language behavior is related to it.

Logan, Beier, and Kincaid (1956) maintain that resistance to extinction is a direct function of the degree to which the mediating response persists beyond the time at which reinforcement usually occurs. "Postreinforcement time cues" do not occur during 100% acquisition because reinforcement is given every time as soon as *S* enters the goal box. They do, however, occur during extinction, evoking the mediating response and quick extinction results. With longer delays of reinforcement during acquisition such as occur in a varied delay procedure, the occurrence of mediating responses to postreinforcement time cues will be increased because the mediating response is eventually reinforced after the delay. Thus the mediating response continues over a longer time for the partial delay *S*s, and gets rewarded for being prolonged. It will occur over a longer period during extinction also. Because the mediating response lasts longer, and is conditioned, presumably, to the instrumental response, the instrumental response will also last longer. Logan, Beier, and Kincaid (1956) are not entirely clear on this point, however.

Expectancy. Expectancy "theory" was brought into partial reinforcement by Humphreys (1939). He considered that his early partial reinforcement studies were contrary to a drive reduction point of view, but that the data could be handled nicely by a concept of expectancy. Basically, he argued that partial reinforcement resulted in an expectancy of irregular reinforcement and that continuous reinforcement resulted in an expectancy of regular reinforcement. He further stated that it was easier to change from a regular expectancy of one kind (that rewards occur on every trial) to a regular expectancy of another kind (that rewards do not occur on any trial), than it was to change from an irregular expectancy (that rewards occur on only some of the trials) to a regular expectancy.

Lewis and Duncan (1957) attempted to test Humphreys' notion by having their *S*s state, before each trial, the confidence they held that they would win or lose on the next trial. They found no evidence in their data for Humphrey's notion that those with irregular expectancies of winning would continue longer during extinction than those with

regular expectancies. Insofar as Humphreys meant some over-all kind of expectancy, these data are not pertinent since they are trial-by-trial expectancies.

The two-light Humphreys board has been a standard apparatus for studying expectancies—"yes" and "no" statements—for a number of years. Recently several studies have appeared with this device which have used a shift in stimulus presentation probabilities. In order to respond to the postshift probabilities, presumably the responses to the preshift probabilities must extinguish. Thus this situation seems to afford a means, at least indirectly, of studying expectancies during extinction.

Parducci (1957), with a two-choice betting game, gave separate groups three preshift probabilities—15%, 50%, and 70%. Then all groups were shifted to 70%. He found more complete adjustments to the postshift probability when the magnitude of the shift was greatest. He interpreted his results according to a discrimination theory and counter to Humphreys' regularity of expectancy analysis. He also concluded, although the evidence is not very clear on this point, that the expected *permanence* of shift was more important than the magnitude of shift.

Goodnow and her colleagues (Goodnow, 1955; Goodnow & Pettigrew, 1955; Goodnow & Pettigrew, 1956; Goodnow & Postman, 1955) in a number of studies have hypothesized that guesses would be more likely to approach 100% to the most probable light if the situation were a "chance" one than if it were a "problem solving" one. James and Rotter (1958) also hypothesized that the "chance" and "skill" situations would make a difference to resistance to extinction. They gave one set of instructions to indicate that the task involved chance only, and one set to indicate that the task was skill only. Instructions were combined factorially with 50 and 100% reinforcement. Their hypothesis was confirmed. With skill instructions, the 100% group took longer to extinguish than the 50% group. With chance instructions, just the reverse was true. James and Rotter interpret their results as the effects of symbolic processes of some kind, whereas Goodnow and Postman (1955), in a slightly different situation, maintain that "awareness" is not a necessary part of such probability discriminations.

Perhaps every study cited in this report could be reinterpreted according to an expectancy notion, and that is the main weakness of such a point of view. There seems to be no way of disproving it. Certainly not much recent research has been oriented around expectancies, and this concept is apparently moribund as an explanation of the PRE.

Conclusion

The writer began this review of the literature on partial reinforcement in the hope of arriving at some explanation of the PRE. Unfortunately, he is at least as far away from an explanation now as when he started. Conflicts and contradictions in data contribute their share of confusion, but probably more important has been the absence of the right kind of data. In the data section of this paper an attempt was made to discover parametric laws. Except in very few instances, this was impossible. Not many experimenters seem to be interested in how one variable relates to another along the major range of both variables.

Most experimenters are interested in "theory" testing. As a result we have a large number of two or three group experiments, using a widely different array of apparatus, sometimes combined in factorial designs, telling us that our theoretical notions are largely inadequate, but not telling us a great deal more. To determine a parametric law probably at least five points are needed along each dimension. With fewer than five points, the task of stating the law—describing it by an equation—is relatively trivial. With five or more points, and a small number of constants, the dimension can be described with some precision, assuming the data are reliable and that a dimension is actually present. Certainly curve fitting can be an arbitrary procedure, and certainly it is not all there is to theory, but, also certainly, a theory ought to be about something, and parametric data make a wonderful subject matter. But this is just saying, in a more elaborate way, "more research is needed."

The writer feels no desire at present to carry cudgels for any of the "theories" now available, nor does he have a theory of his own to contribute that he has any confidence in, nor does he think it sporting to take further "pot shots" at the existing theories. The writer remains almost as empty of an understanding of partial reinforcement now as when he began to review the literature, but he still considers the problems in the area to be fascinating.

BIBLIOGRAPHY

AMSEL, A. The role of frustrative nonreward in noncontinuous reward situations. *Psychol. Bull.*, 1958, **55**, 102–119.

BABB, H. Proportional reinforcement of irrelevant stimuli and transfer value. *J. comp. physiol. Psychol.*, 1956, **49**, 586–589.

BIJOU, S. W. Patterns of reinforcement and resistance to extinction in young children. *Child Develop.*, 1957, **28**, 47–54.

BIJOU, S. W. Operant extinction after fixed intertrial schedules with young children. *J. exp. Anal. behav.*, 1958, **1**, 25–30.

BITTERMAN, M. E., FEDDERSON, W. E., & TYLER, D. W. Secondary reinforcement and the discrimination hypothesis. *Amer. J. Psychol.*, 1953, **66**, 456–464.

BRADY, J. V. Extinction of a conditioned "fear" response as a function of reinforcement schedules for competing response. *J. Psychol.*, 1955, **40**, 25–34.

BRAND, H., WOODS, A. J., & SAKODA, J. M. Effects of a random versus pattern instructural act in a contingent partial reinforcement situation. *Psychol. Rep.*, 1957, **3**, 473–479.

BROWN, J. S. The generalization of approach responses as a function of stimulus intensity and the strength of motivation. *J. comp. physiol. Psychol.*, 1947, **37**, 92–98.

BROWN, J. S., & BASS, B. The acquisition and extinction of an instrumental response under constant and variable stimulus conditions. *J. comp. physiol. Psychol.*, 1958, **51**, 499–504.

BULLOCK, D. H., & SMITH, W. C. An effect of repeated conditioning-extinction upon operant strength. *J. exp. Psychol.*, 1953, **46**, 349–352.

BUSS, A. H. Some determinants of rigidity in discrimination-reversal learning. *J. exp. Psychol.*, 1952, **44**, 222–227.

CAPALDI, E. J. The effect of different amounts of alternating partial reinforcement on resistance to extinction. *Amer. J. Psychol.*, 1957, **70**, 451–452.

CAPALDI, E. J. The effect of different amounts of training on the resistance to extinction of different patterns of partially reinforced responses. *J. comp. physiol. Psychol.*, 1958, **51**, 367–371.

CLAYTON, F. L. Secondary reinforcement as a function of reinforcement scheduling. *Psychol. Rep.*, 1956, **2**, 377–380.

COWLES, J. T. Food-tokens as incentives for learning by chimpanzees. *Comp. Psychol. Monogr.*, 1937, **14**, 1–96.

CRUM, J. BROWN, W. L., & BITTERMAN, M. E. The effect of partial and delayed reinforcement on resistance to extinction. *Amer. J. Psychol.*, 1951, **64**, 228–237.

DENNY, M. R. The role of secondary reinforcement in a partial reinforcement learn-

ing situation. *J. exp. Psychol.*, 1946, **36**, 373–389.

DENNY, M. W., WELLS, R. H., & MAATCH, J. L. Resistance to extinction as a function of the discrimination habit established during fixed-ratio reinforcement. *J. exp. Psychol.*, 1957, **53**, 451–456.

ELAM, C. B., & TYLER, D. W. The discrimination hypothesis and cue reversal. *Amer. J. Psychol.*, 1958, **71**, 583–586.

ELAM, C. B., TYLER, D. W. & BITTERMAN, M. E. A further study of secondary reinforcement and the discrimination hypothesis. *J. comp. physiol. Psychol.*, 1954, **47**, 381–384.

FATTU, N., AUBLE, D., & MECH, E. V. Partial reinforcement related to "free" responding in extinction with preschool children. *J. exp. Educ.*, 1955, **23**, 365–368.

FATTU, N., MECH, E. V., & AUBLE, D. Partial reinforcement in a bar pressing situation with pre-school children. *J. genet. Psychol.*, 1955, **87**, 251–255.

FEHRER, ELIZABETH. Effects of reinforcement and of pre- and post-reinforcement delays on learning and extinction. *J. exp. Psychol.*, 1956, **52**, 167–176.

FREIDES, D. Goal-box cues and pattern of reinforcement. *J. exp. Psychol.*, 1957, **53**, 361–371.

GLANZER, M. Curiosity, exploratory drive, and stimulus satiation. *Psychol. Bull.*, 1958, **55**, 302–315.

GOODNOW, JACQUELINE J. Determinants of choice-distributions in two choice situations. *Amer. J. Psychol.*, 1955, **68**, 106–116.

GOODNOW, JACQUELINE J., & PETTIGREW, T. F. Effects of prior patterns of experience upon strategies and learning sets. *J. exp. Psychol.*, 1955, **49**, 381–389.

GOODNOW, JACQUELINE, J., & PETTIGREW, T. F. Some sources of difficulty in solving simple problems. *J. exp. Psychol.*, 1956, **51**, 385–392.

GOODNOW, JACQUELINE J., & POSTMAN, L. Probability learning in a problem-solving situation. *J. exp. Psychol.*, 1955, **49**, 16–22.

GORMEZANO, I., & GRANT, D. A. Progressive ambiguity in the attainment of concepts on the Wisconsin card sorting test. *J. exp. Psychol.*, 1958, **55**, 621–627.

GOSS, A. E., & RABAIOLA, E. J. Response strength in a modified Thorndikian multiple-choice situation as a function of varying proportions of reinforcement. *J. exp. Psychol.*, 1952, **43**, 106–114.

GRANT, D. A., HAKE, H. W., & HORNSETH, J. P. Acquisition and extinction of a verbal conditioned response with differing percentages of reinforcement. *J. exp. Psychol.*, 1951, **42**, 1–5.

GRANT, D. A., HAKE, H. W., RIOPELLE, A. J., & KOSTLAN, A. Effects of repeated pretesting with conditioned stimulus upon extinction of the conditioned eyelid response to light. *Amer. J. Psychol.*, 1951, **14**, 247–251.

GRANT, D. A. HORNSETH, J. P., & HAKE, H. W. The influence of the intertrial interval on the Humphreys' "random reinforcement effect" during the extinction of a verbal response. *J. exp. Psychol.*, 1950, **40**, 609–612.

GRANT, D. A., RIOPELLE, A. J., & HAKE, H. W. Resistance to extinction and the pattern of reinforcement. I. Alternation of reinforcement and the conditioned eyelid response. *J. exp. Psychol.*, 1950, **40**, 53–60.

GRANT, D. A., & SCHIPPER, L. M. The acquisition and extinction of conditioned eyelid responses as a function of the percentage of fixed-ratio random reinforcement. *J. exp. Psychol.*, 1952, **43**, 313–320.

GRANT, D. A., SCHIPPER, L. M. & ROSS, B. M. Effect of intertrial interval during acquisition on extinction of the conditioned eyelid responses following partial reinforcement. *J. exp. Psychol.*, 1952, **44**, 203–210.

GROSSLIGHT, J. H., HALL, J. F., & MURNIN, J. Patterning effect in partial reinforcement. *J. exp. Psychol.*, 1953, **46**, 103–106.

GROSSLIGHT, J. H. HALL, J. F., & SCOTT, W. F. Reinforcement schedules in habit reversal—a confirmation. *J. exp. Psychol.*, 1954, **48**, 173–174.

GROSSLIGHT, J. H., & RADLOW, R. Patterning effect of the nonreinforcement-reinforcement sequence in a discrimination situation. *J. comp. physiol. Psychol.*, 1955, **49**, 543–547.

GROSSLIGHT, J. H., & RADLOW, R. Patterning effect of the nonreinforcement-reinforcement sequence involving a single nonreinforced trial. *J. comp. physiol. Psychol.*, 1957, **50**, 23–25.

HAKE, H. W., & GRANT, D. A. Resistance to extinction and the pattern of reinforcement: II. Effect of successive alternation of blocks of reinforced and unreinforced trials upon the conditioned eyelid response to light. *J. exp. Psychol.*, 1951, **41**, 216–220.

HAKE, H. W., GRANT, D. A., & HORNSETH, J. P. Resistance to extinction and the pattern of reinforcement: III. The effect of trial patterning in verbal conditioning. *J. exp. Psychol.*, 1951, **41**, 221–225.

HIRSCH, J. Learning without awareness and extinction following awareness as a func-

tion of reinforcement. *J. exp. Psychol.*, 1957, **54**, 218–224.

HULL, C. L. *Principles of behavior*. New York: Appleton-Century, 1943.

HULSE, S. H., JR. Amount and percentage of reinforcement and duration of goal confinement in conditioning and extinction. *J. exp. Psychol.*, 1958, **56**, 48–57.

HULSE, S. H., JR., & STANLEY, W. C. Extinction by omission of food as related to partial and secondary reinforcement. *J. exp. Psychol.*, 1956, **52**, 221–227.

HUMPHREYS, L. G. Acquisition and extinction of verbal expectation in a situation analogous to conditioning. *J. exp. Psychol.*, 1939, **25**, 294–301.

ISHIHARA, S. An experimental investigation of partial reinforcement. *Jap. J. Psychol.*, 1954, **25**, 9–100.

JAMES, W. H., & ROTTER, J. B. Partial and 100% reinforcement under conditions of chance and skill. *J. exp. Psychol.*, 1958, **55**, 394–403.

JENKINS, W. O., & RIGBY, M. K. Partial (periodic) versus continuous reinforcement in resistance to extinction. *J. comp. physiol. Psychol.*, 1950, **43**, 30–40.

JENKINS, W. O., & STANLEY, J. C., JR. Partial reinforcement: A review and critique. *Psychol. Bull.*, 1950, **47**, 193–234.

JONES, M. B. An experimental study of extinction. *Psychol. Monogr.*, 1953, **67** (19, Whole No. 369).

KANFER, F. H. The effect of partial reinforcement on acquisition and extinction of a class of verbal responses. *J. exp. Psychol.*, 1954, **48**, 424–432.

KATZ, S. Stimulus aftereffects and the partial reinforcement extinction effect. *J. exp. Psychol.*, 1957, **53**, 167–172.

KENDLER, H. H., & LACHMAN, R. Habit reversal as a function of schedule of reinforcement and drive strength. *J. exp. Psychol.*, 1958, **55**, 584–591.

KENDLER, H., PLISKOFF, S., D'AMATO, M., & KATZ, S. Nonreinforcements vs. reinforcements as variables in the partial reinforcement effects. *J. exp. Psychol.*, 1957. **53**, 269–276.

KINTSCH, W., & WIKE, E. L. Habit reversal as a function of length of partial delay of reinforcement. *Psychol. Rep.*, 1957, **3**, 11–14.

LAUER, D. W., & CARTERETTE, T. S. Changes in response measures over repeated acquisitions and extinctions of a running habit. *J. comp. physiol. Psychol.*, 1957, **50**, 334–338.

LAUER, D. W., & ESTES, W. K. Successive acquisitions and extinctions of a jumping habit in relation to schedule of reinforcement. *J. comp. physiol. Psychol.*, 1955, **48**, 8–13.

LEWIS, D. J. Partial reinforcement in a gambling situation. *J. exp. Psychol.*, 1952, **43**, 447–450.

LEWIS, D. J. Acquisition, extinction, and spontaneous recovery as a function of percentage of reinforcement and intertrial intervals. *J. exp. Psychol.*, 1956, **51**, 45–53.

LEWIS, D. J., & COTTON, J. W. Learning and performance as a function of drive strength during acquisition and extinction. *J. comp. Psychol.*, 1957, **50**, 189–194.

LEWIS, D. J., & COTTON, J. W. Partial reinforcement and nonresponse acquisition. *J. comp. physiol. Psychol.*, 1958, **51**, 251–254.

LEWIS, D. J., & DUNCAN, C. P. The effect of partial reinforcement and length of acquisition-series upon resistance to extinction of a motor and a verbal response. *Amer. J. Psychol.*, 1956, **69**, 644–646. (a)

LEWIS, D. J., & DUNCAN, C. P. Effect of different percentages of money reward on extinction of a lever-pulling response. *J. exp. Psychol.*, 1956, **52**, 23–27. (b)

LEWIS, D. J., & DUNCAN, C. P. Expectation and resistance to extinction of a lever-pulling response as a function of percentage of reinforcement and amount of reward. *J. exp. Psychol.*, 1957, **54**, 115–120.

LEWIS, D. J., & DUNCAN, C. P. Expectation and resistance to extinction of a lever-pulling response as a function of percentage of reinforcement and number of acquisition trials. *J. exp. Psychol.*, 1958, **55**, 121–128. (a)

LEWIS, D. J., & DUNCAN, C. P. Vicarious experience and partial reinforcement. *J. abnorm. soc. Psychol.*, 1958, **57**, 321–326. (b)

LINTON, H. B., & MILLER, N. E. The effect of partial reinforcement on behavior during satiation. *J. comp. physiol. Psychol.*, 1951, **44**, 142–148.

LOGAN, F. A. A comparison of avoidance and non-avoidance eyelid conditioning. *J. exp. Psychol.*, 1951, **42**, 390–393.

LOGAN, F. A., BEIER, E. M., & ELLIS, R. A. Effect of varied reinforcement on speed of locomotion. *J. exp. Psychol.*, 1955, **49**, 260–266.

LOGAN, F. A., BEIER, E. M. & KINCAID, W. D. Extinction following partial and varied reinforcement. *J. exp. Psychol.*, 1956, **52**, 65–70.

LONGNECKER, E. D., KRAUSKOPF, J., & BITTERMAN, M. E. Extinction following alternating and random partial reinforcement. *Amer. J. Psychol.*, 1952, **65**, 580–587.

McCLELLAND, D. C., & McGOWN, D. R. The effect of variable food reinforcement on the strength of a secondary reward. *J. comp. physiol. Psychol.*, 1953, **46**, 80–86.

MacKINTOSH, I. The resistance to extinction of responses acquired under irregular conditions of learning. *J. comp. physiol. Psychol.*, 1955, **48**, 363–370.

McNAMARA, H. J., & WIKE, E. L. The effects of irregular learning conditions upon rate and permanence of learning. *J. comp. physiol. Psychol.*, 1958, **51**, 363–366.

MARX, M. H. Resistance to extinction as a function of continuous or intermittent presentation of a training cue. *J. exp. Psychol.*, 1958, **56**, 251–255.

MASON, D. J. The relation of secondary reinforcement to partial reinforcement. *J. comp. physiol. Psychol.*, 1957, **50**, 264–268.

MOWRER, O. H., & JONES, H. M. Habit strength as a function of the pattern of reinforcement. *J. exp. Psychol.*, 1945, **43**, 293–311.

NOTTERMAN, J. M. A study of some relations among aperiodic reinforcement, discrimination training, and secondary reinforcement. *J. exp. Psychol.*, 1951, **41**, 161–169.

NOTTERMAN, J. M., SCHOENFELD, W. N., & BERSH, P. J. Partial reinforcement and conditioned heart rate response in human subjects. *Science*, 1952, **115**, 77–79.

PARDUCCI, A. Alternative measures for the discrimination of shift in reinforcement-ratio. *Amer. J. Psychol.*, 1957, **70**, 194–202.

PERKINS, C. C., JR., & CACIOPPO, A. J The effect of intermittent reinforcement on the change in extinction rate following successive reconditionings. *J. exp. Psychol.*, 1950, **40**, 794–801.

PETERSON, L. P. Variable delayed reinforcement. *J. comp. physiol. Psychol.*, 1956, **49**, 232–234.

PROKASY, W. F., JR. The acquisition of observing responses in the absence of differential external reinforcement. *J. comp. physiol. Psychol.*, 1956, **49**, 131–134.

RAZRAN, G. Partial reinforcement of salivary CRs in adult human subjects: Preliminary study. *Psychol. Rep.*, 1945, **1**, 409–416.

REYNOLDS, W. F. Acquisition and extinction of the conditioned eyelid response following partial and continuous reinforcement. *J. exp. Psychol.*, 1958, **55**, 335–341.

RUBIN, L. S. A demonstration of superior resistance to extinction following continuous reinforcement as compared with partial reinforcement. *J. comp. physiol. Psychol.*, 1953, **46**, 28–32.

SALTZMAN, I. J. Maze learning in the absence of primary reinforcement: A study of secondary reinforcement. *J. comp. physiol. Psychol.*, 1949, **42**, 161–173.

SCOTT, E. D., & WIKE, E. L. The effect of partially delayed reinforcement and trial distribution on the extinction of an instrumental response. *Amer. J. Psychol.*, 1956, **69**, 264–268.

SHEFFIELD, F. D., ROBY, T. B., & CAMPBELL, B. A. Drive reduction versus consummatory behavior as determinants of reinforcement. *J. comp. physiol. Psychol.*, 1954, **47**, 349–354.

SHEFFIELD, F. D., & TEMMER, H. W. Relative resistance to extinction of escape training and avoidance training. *J. exp. Psychol.*, 1950, **40**, 287–298.

SHEFFIELD, VIRGINIA, F. Extinction as a function of partial reinforcement and distribution of practice. *J. exp. Psychol.*, 1949, **39**, 511–526.

SPIVOK, M., & PAPAJOHN, J. The effect of the schedule of reinforcement on operant conditioning of a verbal response in the autokinetic situation. *J. abnorm. soc. Psychol.*, 1957, **54**, 213–217.

STANLEY, W. C., & CLAYTON, F. L. An experimental test of Weinstock's hypothesis concerning the effect of partial reinforcement and extinction. *Psychol. Rep.*, 1955, **1**, 421–424.

STEVENSON, H. W., & PIROJNIKOFF, L. A. Discrimination learning as a function of pretraining reinforcement schedules. *J. exp. Psychol.*, 1958, **56**, 41–44.

TYLER, D. W. Extinction following partial reinforcement with control of stimulus-generalization and secondary reinforcement. *Amer. J. Psychol.*, 1956, **69**, 359–368.

TYLER, D. W., WORTZ, E. C., & BITTERMAN, M. E. The effect of random and alternating partial reinforcement on resistance to extinction in the rat. *Amer. J. Psychol.*, 1953, **66**, 57–65.

WEINSTOCK, S. Resistance to extinction of a running response following partial reinforcement under widely spaced trials. *J. comp. physiol. Psychol.*, 1954, **47**, 318–322.

WEINSTOCK, S. Acquisition and extinction of partially reinforced running response at a 24-hr. intertrial interval. *J. exp. Psychol.*, 1958, **56**, 151–158.

WICKENS, D. D., SCHRODER, H. M., & SNIDE, J. D. Primary stimulus generalization of the GSR under two conditions. *J. exp. Psychol.*, 1954, **47**, 52–56.

WICKENS, D. D., & SNIDE, J. P. The influence of a component of a complex stimulus on resistance to extinction of the complex itself. *J. exp. Psychol.*, 1955, **49**, 257–259.

WIKE, E. L. Extinction of a partially and

continuously reinforced response with and without a rewarded alternative. *J. exp. Psychol.*, 1953, **46**, 255–256.

WIKE, E. L., & MCNAMARA, H. J. The effects of percentage of partially delayed reinforcement on the acquisition and extinction of an instrumental response. *J. comp. physiol. Psychol.*, 1957, **50**, 348–351.

WILSON, M. P. Periodic reinforcement interval and number of periodic reinforcements as parameters of response strength. *J. comp. physiol. Psychol.*, 1954, **47**, 51–56.

WILSON, W., WEISS, E. J., & AMSEL, A. Two tests of the Sheffield hypothesis concerning resistance to extinction, partial reinforcement and distribution of practice. *J. exp. Psychol.*, 1955, **50**, 51–60.

WYNNE, L. E., & SOLOMON, R. L. Traumatic avoidance learning: Acquisition and extinction in dogs deprived of normal peripheral autonomic function. *Genet. psychol. Monogr.*, 1955, **52**, 241–284.

ZIMMERMAN, DONALD W. Durable secondary reinforcement: Method and theory. *Psychol. Rev.*, 1957, **64**, 373–383.

(Received February 2, 1959)

PSYCHOLOGICAL BULLETIN
Vol. 57, No. 1, 1960.

"LAND! LAND!"

GORDON L. WALLS
School of Optometry, University of California, Berkeley

On every hand, one hears the cry that a new continent has been discovered in the realm of visual science. Its supposed magnitude is indicated by the titles of some of the articles written about it—"A Bombshell in Color Theory"; "An Astonishing New Theory of Color"; "New Light on the Eye"; and so on.

The discoverer is Edwin H. Land, originator of Polaroid and the one-minute camera, and an undoubted inventive genius in the Edison-Burbank-Kettering tradition. The claims and predictions made by him and for him are not based upon mere novelties of technology, new optical tricks, new types of filters or of photographic dyes. The underlying discoveries are supposedly so very basic that "a theory of color that has stood for nearly 300 years has suddenly been overthrown . . . the eye does not need red wave lengths of light to see red, does not need orange to see orange (etc.)"; for "scientists since Newton have been completely fooled about the way the eye sees color" so that "every textbook dealing with color will have to be rewritten" (Bello, 1959). We are told that "the recent phenomenon of Edwin Land's essentially 17th-century type of discovery in the field of color vision . . . really shakes every scientist . . ." (Weaver, 1959).

Land's findings are said to be

> rightly jarring the imagination of all workers in the field of color reproduction. His public demonstrations made with two slide projectors and usually only one color filter have achieved multicolor effects on the screen which sometimes could not be told from Ektachromes. . . . Land's goal, of course, is a practical, economical one-step method of achieving full-color prints as simply as black-and-white pictures are now made in Polaroid cameras.[1] The applications, however, could be much more widespread and commercially important. Besides motion pictures and television, which are closely related to the present projection picture tests, a perfection of a similar system for reflection prints might . . . simplify color ink-on-paper systems" (Cros, 1959).

Again, while Land's discoveries

> are not yet ready to be applied to color photography [still] there seems little question that this new knowledge will eventually lead to new color-reproduction techniques that will make the present methods look primitive and clumsy. Television especially should benefit from the knowledge that the eye can utilize an entirely different sort of information from what had formerly been thought necessary to produce color (Bello, 1959, p. 205).

It will be noted that none of the above quotations is attributed to Land. What he himself has said in print is more conservative. The whole presentation of his material was planned for the *Proceedings of the National Academy of Sciences*. Only Parts I and II have so far appeared (Land: 1959a, 1959b), but they are perfectly dignified. His *Scientific American* article (Land, 1959c) is in a more popular vein, but Land says that wherever it is more excited or seems to claim more than his other two articles, this represents editorial change over which he had no control. He did not even know that the paper of Cros (1959) was in preparation, and he disclaims responsibility for the extravagant article of Bello (1959).

The story goes that early in 1955

[1] Land is developing color material for the one-step camera, but it makes no use of the color-visual principles Land thinks his projection phenomena have uncovered.—G.L.W.

Land did begin to consider the possibility of devising a quick process for the Polaroid camera which would yield a print in color in a single while-you-wait operation. To begin to inform himself at firsthand of the problems that would be involved, Land set up, in three projectors, orthodox color-separation transparencies which had been prepared photographically through conventional red, green, and blue gelatine "taking" filters. Screened in registry through similar filters, the images built up additively a full-color reproduction of the original camera scene.

While trying various filters, beam intensities, and so on, Land happened to note that when the blue-filtered projector was turned off there was only a little deterioration of the quality of the reproduction on the screen. In a further "accident," Land removed the filter from the green projector. The intensified white light from it made the picture quite unsatisfactory. But Land's collaboratrix, Meroë Morse (who for a decade has been Land's right arm), noticed that on the screen there were still traces of hues which seemingly should not have been there. At the moment, this was dismissed as a "fatigue" (afterimage) effect. Late that night, however, Land returned to the setup and dimmed down the white projector. "Instantly, the scene burst into lifelike color" (Bello, 1959); and since with indefinitely long observation the color array persisted, it was clearly no afterimage phenomenon.

From then on Land's copy, whether still life or animate, was regularly photographed through only two filters. A positive transparency made from a black-and-white negative taken through a red filter (passing only the longer wave lengths) constituted a "long record" and was ordinarily projected through a red filter. A negative simultaneously made through a broad-band green filter provided a black-and-white positive "short record" of the copy. This was rather a *medium* record, considering what wave lengths and energies would have come through a blue taking filter. The second projector, in which it was placed, usually had no filter over its front lens.

Any object in the original scene was now represented on the screen by a patch of red light coming through a neutral area of a particular density in the long record, and by a congruent patch of white light sent through an identically shaped and sized area of the short record, having (except by chance) a higher or a lower neutral density. The screen "object" was a particular ratio-and-sum of red- and white-light luminances.

An elaborate picture made up of just such objects "should" have presented them all as reds, pinks, whites, and blacks, depending upon the relative and absolute densities of the corresponding patches of the long- and short-record transparency slides: mixtures of red and white lights in the dark-surrounded 3° field of such an instrument as a tricolorimeter could not yield orange, yellow, green, blue, or purple. Land's screen images did show all these hues and, moreover, in "correct hierarchical order"; that is, warm-colored objects of the copy were rendered by about the same warm colors on the screen, and cool-colored objects were represented by cool colors in the picture. Skin looked like skin—in fact more so than in many Ektachrome and Kodachrome pictures.

To Land's mind, here was a great mystery. Where were all the extra

colors coming from, beyond those his projection lights could mix to give, when classical color theory could not explain them? By "classical color theory" Land does not mean isolated color-vision theory (Young-Helmholtzian, Heringian, or other), but this together with the relations of hues to wave lengths in the white-light spectrum, Grassmann's laws of color mixture, and particularly the fact that to obtain all hues and white, three additive primary lights are considered to be necessary and sufficient.

If the light coming to the eye from a patch of screen was seen as having a color which a colorimetrist would not expect to see in a mixture of those two lights at those intensities, then classical color theory was inadequate and outmoded. Some entirely new theory had to be developed to account for all the unexpected colors which one obtained by two-light, one-color projection on a screen (and perhaps in no other way).

Quite soon (May 1955) Land gave the first of his public lectures with demonstrations. The matter then went back into the laboratory, to emerge again in 1957. In that year and the next, Land made presentations to various scientific meetings and university groups, at Washington, Detroit, Rochester, Harvard, and Columbia. The order in which the phenomena came to be shown was as outlined in Land's first publication (1959a), where they are described as numbered experiments.

In an orderly way these experiments bring out that: either the red picture or the white picture is monotone or achromatic, but the registry of both makes a tremendous difference since full color then appears; the picture reappears in full color instantly when a flood of room light, which has washed out the screen image, is turned off; enough room illumination can be permitted to enable the perception of colors apart from the screen, and will not interfere with the colors on the screen; changing the subtense of an "object" at the observer's eye does not change its color; interchanging the long and short records turns red hair and lips greenish or blue-green, showing that the colors are not memory colors nor expectation colors; the hues seen are not much changed by large alterations of beam intensity ratios nor by doubling the contrasts in the short record with superposed duplicates; only red, pinks, and white are seen if the long-record slide is removed from the red-filtered projector—although many intensity ratios are still there; and so on.

Land's oral presentations in Eastern cities never came strongly to the awareness of us who stay on the West Coast. I first learned of his projection color phenomena when a colleague, who travels much more, pressed Part I (Land, 1959a) upon me with the solemn assurance that this was the most important development in color vision in the past 50 years. I read it and forgot it, for I could not see that anything more than "simultaneous color contrast" was involved in Land's securing of hues from red to blue-green when his projection lights were red and green, or red and tungsten "white." Much later, Land's popular article (Land, 1959c) was forced upon me, when I was requested to deal with it as a section editor for a condensation journal. By this time Part II (Land, 1959b) had appeared, as also the articles in *Fortune*, *Graphic Arts Monthly*, and *Architectural Forum* (Anon., 1959), which were stirring up great excitement and wonderment among

all sorts of both scientific and non-scientific people.[2]

In print and in person, Land makes about a dozen serious claims. They cannot be dealt with in either the direct or the reverse order of their "importance," for lesser ones have to be explained before greater ones can be appreciated and vice versa. They will be taken up here, therefore, in more or less an order of comprehensibility. First, however, there is to be considered the question of novelty of fact.

Land no longer makes a claim as regards the novelty of the rich visual results of his simple means. In 1955–1957 he learned that he was by no means the first to find that two color-separation transparencies are much more than two-thirds as good as three, and that when they are added, one color filter is much more than half as good as two. His latest article (Land, 1959c) gives no hint that he knew this (perhaps, one of the editor's doings?), but in his first paper (Land, 1959a) a footnote admitted it. As early as 1914, motion picture systems were patented in which alternate frames were taken through a red filter and no filter or through red and green filters, and projected through red and green ones or through a red alternating with none.

These systems, and various others which used no filters, but two-dye films which dodged the great difficulties of making three-dye-primary films, gave excellent results so long as the cameraman could dictate the colors of the copy. Pure blues had to be foregone; but there are so few such colors in outdoor nature that they were easily avoided, and the faithfulness of color reproduction was satisfactory. If flesh tones, *i.e.* skin colors, were good then everything else was good enough. It actually seemed a waste to use a three-color process for a western movie, or even for indoor stories where furniture and costume colors could be selected to avoid those which would come out "wrong."

Even Land's own favorite method for still pictures is not new. In a memorable Optical Society meeting in New York on March 5, 1943, I witnessed demonstrations by Ralph Evans of Eastman Kodak Company, in one of which he projected long and short records taken through the same filters Land now uses, and with red and yellow filters on the two projectors. The hues on the screen, according to notes made at the time, included red, orange, yellow, green, purple, and magenta. Unwittingly, Land has even used subject matter similar to one of Evans's own favorites: a redheaded girl wearing a green sweater. The two techniques could hardly be closer, for Land was then using red and orange filters; and even when he is using no filter on one projector and calls the light from it white, this is the strongly yellowish light coming directly from a tungsten filament.

With long and short records of elaborate still-life copy such as as-

[2] It appeared that I should digest these publications and write, if anything, a commentary covering all of them. I wrote to Land to ask whether there were still other articles which I should embrace. The upshot was that I was invited to Cambridge to be shown the phenomena on their home grounds. I might then find myself unchanged in opinion, or disarmed, or even converted. I am grateful for the experience (early in July, 1959), and must say that Land and his devoted associates were most kind, generous with their time and facilities, and patient and fair in argument.

Statements here attributed to Land without literature citation are to be considered in the category of "personal communications," transmitted during my visit to his laboratory.

semblages of groceries or bric-a-brac, projected with red and green beams or with red and yellow ("white") ones, it is clearly not the results on the screen which are novel, but Land's explanations of them. Land's full-color results are only the latest of a series of rediscoveries of the same phenomena—first seen, in two-primary projected photographs, at least as early as 1897.

A claim, as good as any to commence with, is that it is just because the subject matter is elaborate, and rendered by good photography, that the many colors appear at all. What is on the screen, Land calls a total image, a complete image, a natural-image situation. The colors seen, except for those of the projection lights and additive mixtures thereof, do not appear when the same lights are mixed in a simple isolated spot. Classical color-vision theories, classical color-mixture laws, do not explain or predict the extra colors. In any place on the screen, the wave lengths and energies just there do not determine the color seen there. If changing the wave lengths changes the colors, it does not do so according to the rules of additive color mixture. The rays of different wave length are not themselves color-making; instead, colors come from the interplay of longer and shorter wave lengths over the entire scene, and it does not much matter what those wave lengths are so long as some are longer and some are shorter. The colors of monochromatic lights and those of colorimeter mixtures are all one big special case, and everyone since Newton has been misled by it as regards how colors arise in "natural" situations.

It is easy to demonstrate that it is Land's color array which is the special case. If his screen is scanned through a slender tube, only spots of red, pink, and white can be found; all other colors vanish. One can take the same tube around the lighted room or outdoors, and view anything and everything, but hardly any objects will change color hardly at all when seen thus. Land's screen pictures are most *un*natural images.

Now it is perfectly true that the wave lengths and energies at a screen locus do not determine the color seen there. But this does not mean that the color is independent of the light at that locus. It is utterly dependent upon the light in the "object" in relation to the light *around* the object. Even in situations much simpler physically, it is easy enough to see a spot as having any of several hues when spot and surround are constituted of nonselective materials and illuminated by a single monochromatic light. The relative intensities created by relative reflectances are governing here, as shown by Helson (1938) and by Judd (1940), and expressed in Helson's principle of "color conversion" with respect to adaptation levels. In Land's material, one ratio-and-sum of two projection lights represents the object. Other ratios-and-sums surround it—if they were not "other," there would be no object, for there would be no contour.

In his own kind of quantitative analyses, Land has paid much attention to the ratio of the intensity of each kind of light in each object to the maximal intensity of that same light (at some place on the screen). He thus takes careful note of the relative intensity of each light in the object, and of the intensities of the two object-lights relative to each other. He completely neglects the other absolute and relative intensities of the same two lights which are

hitting the screen around the object's contours. There are spot-and-surround situations all over Land's screen, and the color of any spot is at the mercy of a *different* spectral composition in its surround unless the luminance of the spot is much higher than that of the surround. No wonder, then, that the color seen in any spot seems to be "independent" of the physics of what is in the spot. To a colorimetrist, Land's most prized phenomena would constitute intolerable artifacts. Colorimetry must be done in a dark or at least a neutral surround. If a sample were matched by mixing in a colored surround of uncontrolled chromaticity, the specification found for the sample would be quite meaningless.

The explanation of his extra colors which Land obviously must controvert, before he can expect our attention to any other explanation, is one in terms of simultaneous color contrast. The expression I myself prefer is "spatial induction (of complementary hue)" for which I shall use induction for short, here. Still a third term for the same phenomenon is "lateral adaptation" which Evans (1943, 1948) points to, along with "general" and "local" adaptation, as being involved in the viewing of any sort of picture in color. For the special case where both spot and surround have the same chromaticity but different intensities, Helson's classical "color conversion" is at work.

There are certain facts about induction which are not generally known, but which are involved in Land's situations (with their action concealed there), and which can be brought out also in the simplest spot-and-surround situations. One usually thinks of induction as it is shown off in an elementary laboratory course, with a square inch of gray paper laid upon a large sheet of colored paper. Here, the phenomenon seems tame indeed. The complementary coloration induced into the gray is not very convincing, although it becomes more so if a sheet of tissue is laid over all. Tongue in cheek, one tells students that this blurs the contour, and that this facilitates induction across it.

The induction product is far more vivid when it is obtained on a screen with a broad annulus of colored light surrounding a disc of white light from a second projector. At an optimal intensity of the spot, its saturation is maximal. Here again, the effect is promoted if the mutual contour of surround and spot is deliberately destroyed, by moving either slide so that disc and annulus are a bit out of registry.

If now a flood of dim white light is put over the screen with a third projector and gradually increased in intensity, one finds that the colored annulus is quickly washed out—a newcomer entering the room would not dream that any colored light was there; but, the colored spot is as saturated as ever, Kirschmann's laws to the contrary. Specifically, if the annulus is blue the spot is yellow, and when the white wash has completely desaturated the blue the spot still glows like a sun. The durability of the induced color has to be seen to be believed—the white wash cannot wash *it* out. Here is a color which has no visible means of support, and is as unexpected as any color in one of Land's pictures; for, a spectrophotometer will show only white light in the spot and the eye sees only white surrounding it.

When such colors are induced into the very real-looking objects which even a black-and-white transparency

puts on a projection screen,[3] their vividness—their apparent saturation, if the term is allowable—is tremendous.

Whatever hue is induced into a spot is inevitably complementary to whatever light, in the surround, is doing the inducing. This is not to say, however, that the surround and spot will necessarily be seen in complementary colors. In Land's pictures, this is rarely the case—which unquestionably influenced him to reject a "contrast" explanation. In the first place, complementariness here is not in the sense of mixture complements—those paired lights which, mixed in a particular intensitive ratio, will comprise a white light—but in the sense of a colored light and the color of its after-image. The mixture complement of a light and its induction or after-image complement are not just the same in hue; there are differences, for known reasons (Helson, 1938; Judd, 1940; Wilson and Brocklebank, 1955). In the second place, a physically chromatic surround may not appear colored at all; or the color seen in a spot may be "mixed" from the direct color of the light actually there and another color induced into it. When a surround is mixed from two beams of light they may both be colored, or, e.g., red and tungsten "white," and in either case the mixture will tend to be seen as white in consequence of general chromatic adaptation. But the red-light content acts as if it were alone, in causing lateral adaptation of the retinal area subtending the spot. Truly white light in the spot would now appear white-minus-red, i.e., blue-green. But if the spot light is itself a little reddish, some of its direct redness can combine with the induced greenness and a little of the blueness to give a basis for neutrality, and the remaining direct redness and induced blueness will make the spot purple. If on the other hand the spot light is yellowish, this direct huedness may kill the blueness of the hybrid induced hue and allow the spot to appear purely green. Furthermore, if a spot is substantially brighter than its surround it can exhibit the hue of either projection light or any additive mixture of the two, while in a spot much dimmer than its surround, induced *blackness* can create maroon, navy blue, olive, brown, and other "synthetic" surface colors. All of these possibilities give way to others, with any manipulation of the ratio and sum of the two kinds of light blended in the surround.

Land's extra colors have a kinship, stronger than he admits, with colored shadows. Produced on a screen with two projectors and a filter or two, shadow colors always come in pairs,[4] which have interesting properties. If both projection lights are colored but not complementary, then quite regardless of what their colors are, the hue of neither shadow is ever the hue of either light since the surround

[3] An almost forgotten word for a slide projector is *stereopticon* (most likely to be used, nowadays, by mistake for *stereoscope*). It derives from the corporeality which images have when one sees them on a textureless screen and is not forced to see them in a plane, as on a paper print.

[4] The most familiar of all colored shadows, the blue ones in a snowscape, may appear to be exceptional. When objects are illuminated both by yellow sunlight and white cloud-light, blue is seen wherever an object blocks the sunlight from the snow. The unseen (adapted-out) yellow surrounding the shadow induces complementary blueness into the white light which, alone, occupies the shadow. No yellow mate to the blue shadow is seen, because the huge white cloud source creates no shadows at all.

of each shadow is not purely either light. But, the hues of the shadows are invariably complementary to each other. This fact has been known so long that its first discoverer is lost —while its most recent investigator (H. Self, a student of Helson's) has not yet published his work. If the light from one projector is truly white and the other light is hued—say, green—we have a special case. The hue of one shadow will be that same green since the surrounding white-plus-green can induce nothing that will alter it; the hue of the other shadow will be red-purple, induced into the white light in the shadow by the impure green light outside the shadow. Here, side by side, are a "purely physical" (green) shadow and a "physiological-psychological" (magenta) one. It is important to note that both shadow colors can be highly and equally saturated while the rest of the screen is acceptably white and necessarily brighter than either shadow—again Kirschmann's Laws to the contrary.

With Land-type long- and short-record slides in red and white projectors, one has only to displace one slide to throw the two records out of registry on the screen. One then sees that the "objects" in one slide are represented by a set of colored shadows, not absolute but partial to different extents, but all of the same hue and differing only in saturation and lightness. The other slide is also putting up a set of shadows, of one other (and complementary) hue; and of course the two hues are salmon and aquamarine. Upon pushing the displaced slide back into registry, one can think—for a time—that the additional colors are all created simply by superimposing two complementary sets of colored shadows upon each other.

When Land claims to be getting full color, this is an exaggeration in two respects. The term implies a faithfulness of reproduction of the copy colors greater than Land does get. It further implies that all possible hues are there, or obtainable, even if some objects have slightly wrong hues. The composition on the screen does exhibit many, many colors (a color being a hue at a given saturation and a particular lightness). If the observer penetrates the glamor of the multiplicity of saturations and lightnesses creating "tints" and "shades," and beautiful surface colors (present by grace of induced blackness), and resolutely hews to the hue circle, he finds that there are gaps in it. Nothing on the screen is really blue without admixture of greenness, and nothing is green without contamination with blueness. Only in casual viewing can one think that some of the palest blue-greens are blue; some of the darkest, green. Land had considerable trouble making a lemon *look* like a lemon with red and white projection—he told me that it took three months of trial and error with taking filters, exposures, and darkroom work.

What can be expected from various ratios-and-sums of red and yellow illuminations inside and outside of object contours? Red, pink, and white of course, with the whites being seen wherever sums are high and the ratio of the two lights is the same as their integrated ratio over the screen (establishing the observer's "white" as his general chromatic adaptation level). Wherever about this same ratio occurs but with a lower or very low sum, there will be a gray or an acceptable black. Upset the ratio but keep the sum fairly high, and there can be oranges and yellows. Then, there will be blue-greens induced by a strong red-light content of surrounds, even when these are

adapted out and appear white, or exhibit induction colors themselves. Some of the blue-greens will be greener, passing for green; some bluer, passing for blue. There will be purples, and blackness-containing synthetic surface colors—tans, browns, maroons, navy, drab, etc. What more could one want? What thunderstruck viewer of this marvel would think to ask whether anything in the copy had been sky-blue? Is it just a coincidence that Land's demonstration pictures have all been made indoors, and show objects *he* selected?

Now for Land to say that none of this is described, explained, or predicted by "classical theory" is to be a bit harsh toward centuries of a complex science and a legion of color scientists living and dead. As old as any branch of color science is the knowledge of induced hues—at low saturations, to be sure, prior to electric sources and projectors. They were being exploited in the design of Gobelin tapestries in Thomas Young's lifetime. No one knows how long ago mothers began to teach their daughters what colors to avoid in hats and scarves, so as not to affect their facial complexions adversely. "Accidental colors" were being monographed when people were still arguing over whether three primaries were necessary and sufficient in painting and printing.

It may be that none of Land's technological predecessors, with two-beam, one-filter color systems, knew just why their systems worked so well. It may be that no one before Evans and Land had the photographic facilities, the filters, and the projectors with which to take such full advantage of induction to produce such breath-takingly beautiful pictures with such saturations of the unexpected colors. But what needs to be explained is the long list of hues that are there and the short list of those that are not there, and this can be done without making use of any part of Land's explanation. One needs only old information and old language. In fact several languages are available: that of simultaneous contrast, that of spatial induction, Evans's three kinds of chromatic adaptation (1943, 1948), Helson's principle of color conversion (1938, 1955), and so on.

Land explicitly denies that any sort of chromatic adaptation is responsible for his color arrays: "The colors . . . appear immediately, and do not alter appreciably with time, and we do not seek to explain them as effects of adaptation. To the best of our knowledge the gamut of color is much larger than that . . . predicted by any theory concerning chromatic adaptation" (Land, 1959a). He formerly demonstrated the immediacy of appearance of the colors by turning on strong room lights which obliterated the screen image, and then switching them off to show that "no visible time lapse" occurs before the full-color effect is seen. A newer device, which he demonstrated to me, is a double shutter with which both beams of his dual projector can be flashed to the screen. Even with only a 0.01-sec exposure, the colors are all there. Land considers that since inadequate time is being allowed for any adaptation process, adaptation cannot explain the colors.

Land's trouble here is that in anything he has published, he shows awareness of only the long, slow, intensitive sorts of adaptation which are regulated by the cold-molasses kinetics of photopigment-concentration change. The general and lateral chromatic adaptations which underly contrast phenomena are instant upon presentation of the stimuli (which is

why the contrast is *called* simultaneous). They are more akin to the alpha adaptation of Schouten and Ornstein (1939) than to the photochemical beta adaptation, and they are readily accomplished by the exercise of inhibitions in purely chromatic pathways, leaving *brightness* vision unaltered (see Walls, 1955). No one to my knowledge has had the hardihood to try to measure the tiny time it takes for induced colors to appear. Since the basis of spatial induction is neural and not photochemical, any time consumed is hardly more than is required for a neuronic message to traverse a couple of synapses. As a matter of fact, simple "persistence of vision" could allow a full manifestation of induction effects even if the presentation time of the stimuli were cut below the period consumed in the induction itself.

By rejecting adaptation, Land brings upon himself just as much trouble in explaining whites and grays, as hued areas. The less stray light there is, escaping from projectors etc., the more surely any colored light or some ratio of any two lights, mixed on the screen, will appear white. It is a familiar experience that even in the ruby illumination of a photographic darkroom, a piece of white paper is subjectively white. In Land's projected compositions there need be no broad white backgrounds, and there need be no broad stray-illuminated border formed by empty screen. Whatever two lights he is projecting with, the observer's "white" will be some intensitive ratio of the two, whether one light is colorimetrically white, or even if both are colored. In any case the average ratio over the screen will be what the observer adapts to, and he will see white or gray in any discrete area that happens to contain that ratio. The more heterogeneous the copy may have been chromatically, the more closely the observer's white will approach the ratio coming out of the projectors if the slides are removed from them and the intensity controls (already adjusted for best verisimilitude) are left untouched.

A conviction of Land's which is pregnant with several kinds of significance is that the correct hierarchical order of the colors in his pictures, the similarity of the color of each object to that of its model in the original copy, defies explanation under classical theory or in terms of any sort of chromatic-adaptation phenomena. In effect he says that when he began to use his unorthodox primaries he was not expecting close verisimilitude, and has never strained to attain it, but it is there anyway. It is the fact that he obtains a full series of hues "from red to blue" when using only red and yellow, or yellow and green (etc.) lights, that must seem the most magical aspect of his results, to those who obtain their scientific information from popular magazines. This degree of verisimilitude secured so cheaply, by such simple means, is what piques the cupidity of those who now contemplate lower costs and higher profits in color photography, color television, and color printing, all as a result of Land's research. And, it was this verisimilitude that originally enticed Land far out on his private theoretical limb. Seeing no basis for it in the stimulus situation, Land attributed the fullness of the full-color gamut to a hitherto unsuspected property of the eye itself. His own further tests of his hypothesis that "the eye can build colored worlds of its own" made with ap-

paratus other than simple lantern-slide projectors, have served only to entrench it in his mind.

The fact that in Land's potpourri pictures the right colors end up, for the most part, in the right objects is only made to seem the more remarkable when one considers that all the hues he ever gets may also be obtained on a screen when the slides projected have stemmed from material that never was colored. One may make two montages of pieces of neutral gelatine filters cemented to glass plates, identical excepting that each piece in one montage is replaced in the other by a congruent piece of a different density. Projected through the same filter(s) Land might use, these will yield the same set of hues on the screen. Just as complete "full color" may even be obtained with photographs, when these are of such character that there is no reason to expect any object to appear in any particular color. This is the case with "long" and "short" records photographed through a dead rat with two different wave length bands of X rays (and, of course, no color filters), then projected in registry through color filters which "translate" the radiopacity pattern of the subject matter into a color pattern easier to study (see Mackay & Collins, 1957).

If one could not even speak of "verisimilitude" in colored X-ray translations of heads and legs or in colorized neutral montages, how is it that the hues are not random in Land's pictures also? What makes them click into place in the correct objects, unless Land is right in thinking that the eye itself puts them there because it is viewing a "natural image"? Sadly for Land's hypotheses, the verisimilitude in his pictures is just about as inevitable as it is in Kodachrome. At the same time, paradoxically, it is due to a coincidence.

When Land assembles some objects with hues ranging from red to blue and photographs them through red and green filters, then projects with red and yellowish lights, the long record of a red object is of low density and lets through intense red light, while the short record is dense and lets through little of the other light. Naturally the red object is rendered red. A really green or blue object is represented in the long record by a dense area and in the short record by a pale one. The shorter-waved (or the "white") light predominates in the screen locus of the object. This light however must of necessity appear blue-green or green-blue. This is not at all because the original object had such a color, but because the stronger red illumination around the object-area of the screen induces its complement, and this happens to be an ambiguous hue which satisfies the observer as a reproduction of either a green object or a blue one as the case may be.

It is a beautiful coincidence, at work for Land, that the complement of his red projection light is a blue-green, so close to blue and so very far from red in the spectral series of hues. With selected copy the result is almost full color, and the hasty cannot be blamed for calling it full color. Only pure blue lies beyond the series one does get, for violet is really a purple, the purples are extra-spectral anyway, and moreover Land's method does afford some purples! The reason why pure blue is not obtainable is not that blue wave lengths coming from blue copy pigments are stopped by the green taking filter used to make the short

record; it is that the long-wave projection light employed has a hybrid hue, not a good blue, as its complement. The projection lights, not the copy colors nor the taking filters, are the real determiners of the set of hues the screen can show. Land, however, thinks of them as the least important elements in the process, as we shall see.

If anyone should fail to grasp the fact that the "nearly full" color of Land's pictures and their fair verisimilitude (excellent for long-wave colors; poorer and poorer—but never *too* bad—for shorter-wave ones) are the results of a beautiful and beautifying coincidence, then I would ask: Suppose the physiology of color vision were such that the complement of red was, say, yellow? How good then would be the verisimilitude? '

Land's idea that hue is, or can be, independent of wave length, and his idea that the obtaining of rich coloration from two "primaries" demands a completely new explanation, led him to conclude that "the eye can build colored worlds of its own out of informative materials that have always been supposed to be inherently drab and colorless." When the retina is receiving only two kinds of light which are not very different from each other "the eye" (=the visual system) is apparently able to reassign hue-sensations, which "belong" to the ends of the whole spectrum, to just those two lights, and the intervening hues to mixtures of the two lights.

From a biological point of view, this is by far the most important of Land's claims. It seemingly has profound implications for the course of evolution of human color vision and for its physiology—it seems to demand a complete abandonment of such old-fashioned conceptions as a set of three or four receptor-types each arousing immutably a unique elementary hue-quality irrespective of the wave length(s) of radiation its unique photochemical contents are absorbing.

In order to determine how close together two projection lights illuminating long and short records could be, i.e., how little of the Newtonian spectrum they could embrace and still give his full color, Land first contemplated using interference filters, with much narrower pass bands than gelatine ones have. Too much intensity would be sacrificed for spectral closeness; so, Land had a high-intensity dual monochromator designed and built. With this he could transilluminate either record with any wave length, seeing the two images in registry by means of a 45° half-silvered mirror.

Land now found that with a wide variety of pairs of monochromatic lights, he could secure full color. It has to be remembered that Land uses the word "color" as if it meant hue. With wave lengths too close together the colors were very unsaturated, and when the lights came from certain regions of the spectrum the gamut of hue itself was much restricted. Land has claimed to get a good result, however, even when the two wave lengths were 579 mμ and 599 mμ. The color illustration prepared by *Scientific American* (for Land, 1959c) which purports to show what then appears on the mirror is very hard to believe.

One phenomenon appeared with the dual monochromator which apparently still puzzles Land, although he has promised a discussion of it. This is what he calls "short-wave reversal." When both of the transilluminating wave lengths are short ones, the very few hues he then gets are in the reverse of a correct order; that is, the coolest-colored copy-ob-

jects come out in the least cool hues. This corrects itself if he interchanges the records or the wave lengths so that a violet wave length illuminates the long record and a blue one the short. To my mind the reason for this is that violet has redness in it and blue does not, so that a violet wave length acts like a longer one when it is paired with a blue.

The work with the dual monochromator appeared to support the theory that when only the extremes of a short segment of the whole 300-mμ Newtonian spectrum are available to the retina, the eye exhibits "amazing versatility" and "can respond with a full range of sensation" (i.e., of hues), by reassigning all the hues that belong to a 300-mμ stretch to the wave lengths within a stretch of only 20 mμ or 30 mμ. Somewhere within such a short range there is a point which, Land believes, the eye establishes as a "fulcrum." Wave lengths longer than this, no matter what their absolute values are, are seen with hues which we think of as belonging to the longwave end of the full spectrum, while wave lengths shorter than the balance point are responded to as if they were really short (blue-green, blue) even if they lie in midspectrum.

Out of this reassignment theory arose a corollary hypothesis which Land has quite explicitly outlined (1959c): Suppose the pigments of things around us had very narrow reflection-bands of wave lengths, so that a number of such pigments could be crowded into a short piece of the present visible spectrum, say 30 mμ or 40 mμ long. Suppose also that something should happen to the present solar spectrum reaching the earth's surface—one might imagine light-absorbing materials getting into the atmosphere—so that only that same 30–40 mμ "spectrum" was available to the retina. The colors of the things around us, Land believes, would be just the ones we see now. The reassignment phenomenon prepares us to be transported into any one of many "color worlds."

The basic idea of a "fulcrum" is not wholly wrong, provided one takes the position that it really expresses Land's own unconscious rediscovery of general chromatic adaptation. If one projects pictures or shadows with yellow and blue-green wave lengths and gets red and blue on the screen, the red sensation is not being "assigned to" the yellow wave length but is induced by the blue-green one, and the blue is not assigned to the blue-green but is induced by the yellow. One wave length would indeed appear here to be a fulcrum—the dominant wave length of that Y-BG mixture which, through general chromatic adaptation, had become the observer's "white." But the eye does not select this wave length out of what is coming into it, and use it in a particular way, for it need not be present in either stimulus.

With the original method using slide projectors, Land had found that "white" (unfiltered tungsten-) light could serve as a "primary" just as well as the purest of color-filtered lights. The dual monochromator contained a pair of hinged white surfaces, each of which could be swung into position to occlude one of the diffraction gratings and reflect the undispersed tungsten-lamp light through one slide to the 45° mirror and the observer's eye. This arrangement enabled Land to confirm that when shortwave monochromatic light came through the short record and white light through the long, the picture contained reds as though the white light were a long wave length. When white light was sent through

the short record and long-wave monochromatic light through the long one, the picture exhibited blues as if the white light were a short wave length.

This led Land to pose the question, "what is the wave length which is equivalent to white, the wave length than which white is neither longer nor shorter?" Such wave length would be that which the eye employs as "the fulcrum in the ordinary, sunlit world," with respect to which the visual system assigns a redness sensory quality to the longest other wave length available and a blueness sensation to the shortest other wave length coming into the eye. With white light going through one record and monochromatic light through the other, the observer would find the fulcrum wave length as that one which made the picture go neutral or monotone all over.

All but two out of Land's many subjects, told to watch a red "object" and rotate the wave length control until it started to turn green, found the "fulcrum" at $\lambda 588 \pm 2$ mμ. Considering that the "white" light here was again yellow tungsten-filament radiation, it is immediately evident to anyone with a smattering of colorimetry that this was no experiment, but a simple observation—a determination of the dominant wave length of the tungsten emission, i.e., that wave length which would have the same hue to the observer as the tungsten light itself. No mixture in a surround, of any two lights having the same hue, however different in purity, could possibly induce any hue into any mixture of the same two lights in the surrounded spot. The two odd observers (one male, one female) who found a rather different "sunlight fulcrum" wave length must have had unusual lens pigmentations, unusual luminosity curves, or some such things. Both had normal color vision, for I checked them myself with my own Nagel anomaloscope, which I had with me in Cambridge.

The dual monochromator work made it appear that white light could give a sensation of red, blue, or any color depending upon what monochromatic light it was paired with; and it seemed to show that almost any wave length could arouse a sensation of almost any hue, depending upon what other wave length, far enough away, *it* was paired with (Land, 1959b, 1959c). To demonstrate these things convincingly to a group, with something other than a one-man, eyepiece instrument, Land devised another apparatus in which he took advantage of the fact that one particular wave length of monochromatic light is easily available at very high intensity from an inexpensive source: a standard sodium vapor highway lamp emits $\lambda 589$ mμ.

In one of his two "sodium viewers" (Land, 1959b, 1959c), a short-record transparency with the objects enlarged to natural size is transilluminated by tungsten lamps and viewed through a 45° half-silvered mirror, and the long record (reversed right for left) is transilluminated by a sodium luminaire and seen in the mirror, so that the two color-separation pictures are optically coplanar and congruent. The observer looks into this viewer through a hand-held green filter which transmits $\lambda 589$ mμ, and then has the effect of a dual-projector-and-screen setup with green-filtered and 589 mμ projection lights. In the other viewer the $\lambda 589$ mμ comes through the short record and the long record is tungsten-illuminated through a red filter. In the first viewer, $\lambda 589$ mμ is longer than the dominant wave length of the green short-record filter; in the

second viewer, λ589 mμ is shorter than the dominant wave length of the red long-record filter.

The picture of groceries which Land presents with the green filter contains some objects partly or wholly red, in particular a vivid red can of pepper. Land asserts that here it is the λ589 mμ coming from the pepper-can area which is arousing the red sensation. In the other viewer, where λ589 mμ is the shorter-wave stimulus, some objects appear in cool hues, notably a blue-green book binding. Here, Land claims that λ589 mμ is arousing *that* sensation. So, if one and the same wave length can "be" red at one time and green or blue at another, the thesis that the eye reassigns various hues to wave lengths to suit itself is abundantly proved.

The sodium viewer experiment proves no such thing. Consider what is making the famous pepper-can red through the green filter. The can is dense in the short record, pale in the long, hence is represented by bright 589-light and dim green light. Since the surround of the can in the copy was *not* red, it is represented by relatively intense green light and feeble 589-light. The green induces a strongly red purple into the can area. Here the blueness element of the purple encounters the direct yellowness of the sodium light and they kill each other off (after the fashion of complementation), leaving the redness to be seen alone in the can. The key to what is really going on is given by the individual variation that transpired when I myself viewed this demonstration in Cambridge. For me, the pepper-can was a slightly bluish red, a magenta. I did not develop enough yellowness to kill all the blueness that was there for me. Another person present called the can purely red. A third and fourth called it orangeish-red—for them, there was more than enough yellowness, or else less blueness in the hybrid complement induced by the green light. What was varying here, of course, was the shape of one's personal curve of "the intrinsic saturation of the spectrum."

There are other ways to describe the genesis of the red pepper-can; but they are all in old languages, requiring no supposition that "the eye is revealed to be an instrument of unsuspected and awe-inspiring subtlety." So far as I am concerned, the red of the pepper-can is being seen where the λ589 mμ *is*, but it is the adapted-out green in the can's surround that is putting the red there. The same surround would induce redness into any other light substituted for the λ589 mμ, just so its direct color was low enough in saturation. In the red-filtered sodium viewer, it is just as clear that any green-blue or blue-green seen in any object is only the expected induction product of the more intense red light in the surround, which readily overwhelms the direct yellowness of the λ589 mμ (yellow being the least saturated locus in the spectrum).

Land, believing strongly as he does in his theory that color vision works by assignment of hues to available wave lengths simply according as they are longer or shorter, has been led to a philosophy about the evolution of human color vision which Bello (1959) quotes directly from one of Land's 1957 public lectures: "Would it be possible, in a system which had only a pair of receptors balancing off against each other somehow, for the animal to have a sense of a full range of hues running from red, orange, yellow, green, down into blue? That animal could see the leopard in the leaves in a way that his competitors couldn't. He would survive the leop-

ard . . . and go on to build the world we live in." Elsewhere (1959c) Land has said:

> . . . we have not been describing a two-color theory of vision. . . . It is true, however, that our experiments deal with two packages of information . . . the eye can do almost everything it needs to do with these two packages . . . there is not a very big gap in the sensation scale to be filled . . . the visual process will remain an amazing one from the evolutionary point of view. Why has a system that can work so well with two packages of information evolved to work better with three? . . . we feel that the big jump is obviously from one to two (p. 99).

Translated into language which most color-vision researchers would use, this means that Land views the course of evolution of primate color vision thus: Originally, a single receptor-type afforded only monotone or achromatic imagery. Upon the differentiation of this into two types the system became able to give to consciousness the hues from red to bluish blue-green, some purples, neutrals, surface colors (containing blackness), all by assigning the extremes of the hue series to the extremes of any available stimulatory wave length band. Essentially the only hue missing was purest blue. It is hard to see why the organism went on to add 50% more mechanism (a third receptor-type and all accessory neurons) just to secure this one more hue for the sensory gamut; but this did happen.

The catch here is a double one. In the first place, at a point in evolution prior to the invention of a blueness receptor in the retina, the higher levels of the system would not have embodied mechanism for putting any blueness into any sensation, whether by induction or otherwise. Man's present three-component system affords redness, greenness, and yellowness synthesized from their bases, together with blueness, and all of these hue qualities, either singly or in blends of two, can get into sensations—courtesy of induction—when only red and yellow (say) lights are impinging upon the retina. But these lights, or any two others, or any one light, activate all three components of the system. In an earlier, two-component stage of the system, no more hue qualities than red, possibly (but probably not) yellow, and green could have been enjoyed.

Land (1959c) errs in saying that the central idea of the three-component theory of color vision is that "the eye responds to three different kinds of vibration, and all color sensation is the result of stimulating the three responses in varying degrees of strength." He attributes this to Maxwell and Helmholtz, who instead believed that each receptor-type responds to some extent to every wave length in the visible spectrum. One has to go back to Thomas Young for the short-lived idea that each receptor is activated primarily by one bit of the spectrum (one "kind of vibration") and responds at all to neighboring frequencies only in "forced vibration." Land seems to think that if he uses two projection lights, the hue of neither of which has any blueness about it, he is stimulating only two components and the third might as well never have evolved—all the hues now seen by an observer *with* a three-component system being those which would have been seen by the remote ancestor who did not yet have any blueness component at all.

In the second place, we have people with two-component systems all around us—the dichromatic color-blinds. They comprise three or four very different types, but we know that not one of them sees more than *two* hues in the entire white-light spectrum. The color-normal experi-

ences perhaps 40 hues.[5] Dichromatic vision is "a big jump" above monochromatic, but it does not afford 40 − (pure blue) = 39 hues. Even though the color-normal sees many more colors in a Landian two-projector picture than the simple stimuli seem to be worth, this would not be true for a dichromate. For him, any wave length to one side of a neutral point can induce only the same one-and-only other hue that he gets directly from any wave length on the other side of the neutral point. With long and short records projected through red and green filters, a normal may see on the screen reds, oranges, yellows, chartreuse, greens, blue-greens, purples, and neutrals. A tritanope standing alongside would not see more than one kind of red and one kind of green, each at various saturations down to zero (neutral).

While most of Land's color research thus has to be called much ado about nothing new, it would be a mistake to say that it is all just a flash in the panchromatic! Land's "new coördinate system" and the relationships he can extract from it are probably salvageable (although Judd's 1940 formulations might be as fruitful in the projected-color situations, or more so). It may well be that in the hands of Land and others the coördinate system will enable a quantification and accurate prediction of induction products which never could have been had, or had easily, without it. This may be only my own wishful thinking, stimulated by a considerable sympathy for Land. I myself am not qualified to quantify anything more complex than my bank balance.

Land's coördinates are Cartesian, but nothing else about them is conventional. The ordinate axis is a logarithmic scale of "percentage available long-wavelength stimulus" and the abscissa axis bears a log scale of percentage available short wave length stimulus. With long and short records in projectors or in the dual monochromator, and any two projection lights, any point on the screen finds a place also as a point on a graph with these coördinates. When the beams have been regulated for best color array, the short-record illuminant is cut off and the maximal intensity of the long record image is found and measured, with this luminance then taken as 100%. The relative intensity at any other locus gives an ordinate value for that locus. Abscissa values are found in the isolated short-record image, as percentages of the highest intensity anywhere in the image. One point on the graph may, of course, represent any number of image points where the same absolute luminances of the two kinds of light happen to occur. These various physically identical image loci should not exhibit the same color unless they all have identical surrounds; but Land appears to claim that they do, since on a sample graph, given in all three of his papers, he labels single points with single color names.

Each point on the graph represents a relative intensity relative to another relative intensity, and is supposed to describe and predict a unique color on the screen. This has helped to make it appear to Land that "the eye turns brightness ratios into colors" (Bello, 1959), although, as pointed out above, he considers

[5] This has to be a guess. It is commonly taught that there are about 120 discriminable hues in an equal-brightness spectrum and 20–30 extraspectral purples. There may be 120 values of $\Delta\lambda$, but each is a complex of hue and saturation. The curve of "$\Delta\lambda$ *vs* λ" has never been determined in an *equal-saturation* spectrum. It is not a "hue discrimination curve," but a wave length discrimination curve.

only ratios in spots and neglects ratios in surrounds and ratios of spot ratios to surround ratios. Land emphasizes that his coördinates are dimensionless, scaled in naked numbers. Actually, since a new graph has to be made for every pair of color-separation positives and for every pair of lights used for every picture, the coördinates could be calibrated directly in log luminance. It is hard for me to see why they are not, for the percentage scales do not generalize or universalize the graph. The predictive values of any one graph are confined within itself, for that graph can tell nothing about what colors will be seen, or where, in another picture with other lights or even with the same lights.

Land soon noticed that all of the whites, grays, and blacks on the screen plotted along a 45° line in his coördinate system. This is logical, for on the 45° line the arithmetic ratio of the intensity of one light to the intensity of the other is a constant. If a screen point with this ratio and some sum appears neutral, any other point with the same ratio and another sum should appear neutral; and this ratio is the one which ordinarily would exist all over the screen if the slides should be removed from the apparatus. Land calls the 45° line the "neutral line" or the "gray line"; it could as well be called the "general chromatic-adaptation line" or the "adaptation level." Since warm colors necessarily plot above the 45° line and cool colors below it, one readily sees whence came Land's idea that the Newtonian spectrum is a physical artifact of prisms and gratings and that the real visual scale of sensation runs from warm hues through neutrality to cool hues.

One has to read with caution Land's description of the properties of straight lines other than the 45° one, on a graph in his coördinates. He calls them "achromatic tracks" and includes among these the ordinate and abscissa axes themselves. Just as one has to know that when Land says color he means hue, so also one has to know that when he says achromatic he means achromatic-or-monotone—he makes no distinction between a set of screen points which all have the same hue, and a set which all have no hue. Anything short of full color, he is likely to dismiss as no color.

Among these "achromatic" straight-line loci are the rotated or displaced (but still straight) gray line which results from doubling the contrast in one image (by using two layers of short-record transparency); and a line at right angles to the normal gray line. Such a line is the locus of all that one can get on the screen by projecting any negative in registry with a positive made from it. This is not necessarily an "achromatic" (monotone) situation, for the line may cross the gray line and continue into the "cool" area: although one sees only reds, pinks, and whites if a dark negative goes through a red filter and the bright positive through none, one may get reds, neutrals, and blue-greens if the positive is red-filtered and the negative is tungsten-illuminated.

This negative–positive combination serves to exemplify some situations which puzzle even Land, and baffle me completely. On the screen, each object may be represented by one of the very same ratios of intensity of the projection lights that would be on the screen if long- and short-record positives were in the projectors. But the negative-positive screen picture is monotone or at best bicolor. This is not because the negative is always so dense that all

surrounds contain essentially only the positive projector's kind of light —Land convinced me of that by projecting a negative and positive of his "standard objects slide" (illustrated in Land, 1959a) where, since the background in the copy was a mid gray, it is about equally translucent in both positive and negative.

Another mysteriously "achromatic" situation is obtained by projecting three identical positives made from any negative (color separation, or not) with two superimposed in one projector, the third alone in the other projector, and a color filter over each lens. The screen loci again receive an enormous number of ratios of the two kinds of projection light. Perhaps Land's "doubled gamma" transparencies have always been *too* dense over most of their area, for I have learned of two other people, who have work in progress on this matter, who do get the various hues in the "three identical images" situation.

Thus, there are situations in which brightness ratios as such do not get "turned into colors" by our clever eyes. I don't know why one doesn't always get new colors when one mixture of two colored lights is surrounded by another mixture of the same lights. In all of the situations Land finds to be "achromatic" there is a common feature, in that there are regular progressions of ratios. This in itself is enough to lead him to expect no gaudy results. He justifies his use of elaborately-composed photographs to begin with (instead of "scientific" spots-and-surrounds and montages of "scientific" Munsell papers) by insisting that randomness of the projection light intensity ratios and of their topographic distribution over the screen is essential for the varicolored effect. Such randomized configurations are, he says, natural; and it is in natural images that he has found the truth about color vision.

One of the tests with which he "proves" the necessity of randomness, and the color-killing effect of regularizing in any way the intensity-ratio array on the screen, is the experiment with neutral step wedges. With two such wedges in place of slides, one oriented horizontally in one projector and one vertically in the other, and with a red filter over one projector, one gets on the screen only a huge square which is red in one corner, white in the opposite corner, and pink in between. Yet, a great number of intensity-ratios of the lights are there (256, to be exact), just as in one of Land's beautiful long-and-short-record pictures.

When I asked for black grids to be put on the wedges, which would themselves be in registry on the screen, and this was done, the screen was no longer monotone but exhibited several hues. Any illuminated square now had, nearest to it, other squares in which the ratios were different enough to be capable of inducing something into the first square. To my own satisfaction this showed that it is not randomness per se but *discreteness* that is essential, for the ratios in the squares were still in a regular topographic array. The virtue of Land's photographs of hodgepodges is that the easy way to get discreteness—local changes of ratio large enough to create contours —is by permitting randomness, which is rife in just such photographs.

And with this, I can strike my closing note, which is that I am quite content with being able to explain why Land gets what he does get, using only old knowledge and old language and making no use whatever of Land's fantastic new hypotheses and new concepts. I leave

it for others to explain why he doesn't get what he doesn't get, i.e., why the "unexpected" colors are lacking where he has come to expect them, as in his various "achromatic" situations. In each of these, I am sure, some simple principle—which just happens to escape nonmathematical *me*—will explain the nonappearance of "full color."

REFERENCES

ANONYMOUS. New light on the eye. *Archit. Forum*, 1958, **109**(6), 124–128.

BELLO, F. An astonishing new theory of color. *Fortune*, 1959, **59**(5), 144–206.

CROS, M. A bombshell in color theory. *Graphic Arts Mon.*, 1959, **31**(6), 26–30; 204.

EVANS, R. M. Visual processes and color photography. *J. Opt. Soc. Amer.*, 1943, **33**, 579–614.

EVANS, R. M. *An introduction to color*. New York: Wiley, 1948.

HELSON, H. Fundamental problems in color vision: I. The principle governing changes in hue, saturation, and lightness of nonselective samples in chromatic illumination. *J. exp. Psychol.*, 1938, **23**, 439–476.

HELSON, H. Color and seeing. *Illum. Engin.*, 1955, **50**, 271–278.

JUDD, D. B. Hue saturation and lightness of surface colors with chromatic illumination. *J. Opt. Soc. Amer.*, 1940, **30**, 2–32.

LAND, E. H. Color vision and the natural image. Part I. *Proc. Nat. Acad. Sci.*, 1959, **45**, 115–129. (a)

LAND, E. H. Color vision and the natural image. Part II. *Proc. Nat. Acad. Sci.*, 1959, **45**, 636–644. (b)

LAND, E. H. Experiments in color vision. *Sci. Amer.*, 1959, **200**(5), 84–99. (c)

MACKAY, R. S., & COLLINS, C. C. Color X-ray images and enhanced contrast. *J. Biol. Photog. Ass.*, 1957, **25**, 114–118.

SCHOUTEN, J. F., & ORNSTEIN, L. S. Measurements on direct and indirect adaptation by means of a binocular method. *J. Opt. Soc. Amer.*, 1939, **29**, 168–182.

WALLS, G. L. A branched-pathway schema for the color-vision system and some of the evidence for it. *Amer. J. Ophth.*, 1955, **39**(2, Pt. II), 8–23.

WEAVER, W. Dither. *Science*, 1959, **130**, 301.

WILSON, M. H., & Brocklebank, R. W. Complementary hues of after-images. *J. Opt. Soc. Amer.*, 1955, **45**, 293–299.

(Received July 17, 1959)

PSYCHOLOGICAL BULLETIN
Vol. 57, No. 1, 1960

THE EFFECTS OF VIOLATIONS OF ASSUMPTIONS UNDERLYING THE *t* TEST[1]

C. ALAN BONEAU
Duke University

As psychologists who perform in a research capacity are well aware, psychological data too frequently have an exasperating tendency to manifest themselves in a form which violates one or more of the assumptions underlying the usual statistical tests of significance. Faced with the problem of analyzing such data, the researcher usually attempts to transform them in such a way that the assumptions are tenable, or he may look elsewhere for a statistical test. The latter alternative has become popular because of the proliferation of the so-called nonparametric or distribution-free methods. These techniques quite generally, however, couple their freedom from restricting assumptions with a disdain for much of the information contained within the data. For example, by classifying scores into groups above and below the median one ignores the fact that there are intracategory differences between the individual scores. As a result, tests which make no assumptions about the distribution from which one is sampling will tend not to reject the null hypothesis when it is actually false as often as will those tests which do make assumptions. This lack of power of the nonparametric tests is a decided handicap when, as is frequently the case in psychological research, a modicum of reinforcement in the form of an occasional significant result is required to maintain the research response.

Confronted with this discouraging prospect and a perhaps equally discouraging one of laboriously transforming data, performing related tests, and then perhaps having difficulty in interpreting results, the researcher is often tempted simply to ignore such considerations and go ahead and run a *t* test or analysis of variance. In most cases, he is deterred by the feeling that such a procedure will not solve the problem. If a significant result is forthcoming, is it due to differences between means, or is it due to the violation of assumptions? The latter possibility is usually sufficient to preclude the use of the *t* or *F* test.

It might be suspected that one could finesse the whole problem of untenable assumptions by better planning of the experiment or by a more judicious choice of variables, but this may not always be the case. Let us examine the assumptions more closely. It will be recalled that both the *t* test and the closely related *F* test of analysis of variance are predicated on sampling from a normal distribution. A second assumption required by the derivations is that the variances of the distributions from which the samples have been taken is the same (assumption of homogeneity of variance). Thirdly, it is necessary that scores used in

[1] This project was undertaken while the author was a Public Health Service Research Fellow of the National Institute of Mental Health at Duke University. The computations involved in this study were performed in the Duke University Digital Computing Laboratory which is supported in part by National Science Foundation Grant G-6694. The author wishes to express his appreciation to Thomas M. Gallie, Director of the Laboratory, for his cooperation and assistance.

the test exhibit independent errors. The third assumption is usually not restrictive since the researcher can readily conduct most psychological research so that this requirement is satisfied. The first two assumptions depend for their reasonableness in part upon the vagaries inherent in empirical data and the chance shape of the sampling distribution. Certain situations also arise frequently which tend to produce results having intrinsic non-normality or heterogeneity of variance. For example, early in a paired-associate learning task, before much learning has taken place, the modal number of responses for a group will be close to zero and any deviations will be in an upward direction. The distribution of responses will be skewed and will have a small variance. With a medium number of trials, scores will tend to be spread over the whole possible range with a mode at the center, a more nearly normal distribution than before, but with greater variance. When the task has been learned by most of the group, the distribution will be skewed downward and with smaller variance. In this particular case, one would probably more closely approximate normality and homogeneity in the data by using some other measure, perhaps number of trials for mastery. In many situations this option may not be present.

There is, however, evidence that the ordinary t and F tests are nearly immune to violation of assumptions or can easily be made so if precautions are taken (Pearson, 1931; Bartlett, 1935; Welch, 1937; Daniels, 1938; Quensel, 1947; Gayen, 1950a, 1950b; David & Johnson, 1951; Hornsnell, 1953; Box, 1954a, 1954b; Box & Anderson, 1955). Journeyman psychologists have been apprised of this possibility by Lindquist (1953) who summarizes the results of a study by Norton (1951). Norton's technique was to obtain samples of Fs by means of a random sampling procedure from distributions having the same mean but which violated the assumptions of normality and homogeneity of variance in predetermined fashions. As a measure of the effect of the violations, Norton determined the obtained percentage of sample Fs which exceeded the theoretical 5% and 1% values from the F tables for various conditions. If the null hypothesis is true, and if the assumptions are met, the theoretical values are F values which would be exceeded by chance exactly 5% or 1% of the time. The discrepancy between these expected percentages and the obtained percentages is one useful measure of the effects of the violations.

Norton's results may be summarized briefly as follows: (*a*) When the samples all came from the same population, the shape of the distribution had very little effect on the percentage of F ratios exceeding the theoretical limits. For example, for the 5% level, the percentages exceeding the theoretical limits were 7.83% for a leptokurtic population as one extreme discrepancy and 4.76% for an extremely skewed distribution as another. (*b*) For sampling from populations having the same shape but different variances, or having different shapes but the same variance, there was little effect on the empirical percentage exceeding theoretical limits, the average being between 6.5% and 7.0%. (*c*) For sampling from populations with different shapes and heterogeneous variances, a serious discrepancy between theoretical and obtained percentages occurred in some instances. On the basis of these results, Lindquist (1953 p. 86) concluded that "unless the heterogeneity of either form or vari-

ance is so extreme as to be readily apparent upon inspection of the data, the effect upon the F distribution will probably be negligible."

This conclusion has apparently had surprisingly little effect upon the statistical habits of research workers (or perhaps editors) as is evident from the increasing reliance upon the less powerful nonparametric techniques in published reports. The purpose of this paper is to expound further the invulnerability of the t test and its next of kin the F test to ordinary onslaughts stemming from violation of the assumptions of normality and homogeneity. In part, this will be done by reporting results of a study conducted by the author dealing with the effect on the t test of violation of assumptions. In addition, supporting evidence from a mathematical framework will be used to bolster the argument.

To temper any imputed dogmatism in the foregoing, it should be emphasized that there are certain restrictions which preclude an automatic utilization of the t and F tests without regard for assumptions even when these tests are otherwise applicable. It is apparent, for example, that the violation of the homogeneity of variance assumption is drastically disturbing to the distribution of t's and Fs if the sample sizes are not the same for all groups, a possibility which was not considered in the Norton study. It also seems clear that in cases of extreme violations, one must have a sample size large enough to allow the statistical effects of averaging to come into play. The need for such considerations will be made apparent in the ensuing discussion. There is abundant evidence, however, that both the t and the F tests are much less affected by extreme violations of the assumptions than has been generally realized.

A Sampling Experiment

At this point we will concern ourselves with the statement of the results of a random sampling study. The procedure is one of computing a large number of t values, each based upon samples drawn at random from distributions having specified characteristics, and constructing a frequency distribution of the obtained t's. The present study was performed on the IBM 650 Electronic Computer programmed to perform the necessary operations which can be summarized as follows: (*a*) the generation of a random number, (*b*) the transformation of the random number into a random deviate from the appropriate distribution, (*c*) the successive accumulation of the sums and sums of squares of the random deviates until the appropriate sample size is reached, (*d*) the computation of a t based upon the sums and sums of squares of the two samples, (*e*) the sorting of the t's on the basis of size and sign and the construction of a frequency distribution based upon the sorting operation. The complete sequence of operations was performed internally, the end result, the frequency distribution of 1000 t's, being punched out on IBM cards.

Comments on many of the above operations are relevant and will be made according to their order above.

(*a*) The random numbers consisted of 10 digits, the middle 10 digits of the product of the previously generated random number and of one of a sequence of 10 permutations of the 10 digits (0, 1, 2, · · · , 9) placed as multipliers in the machine. To start the process it was necessary to place in the machine a 10 digit random number selected from a table of random numbers. The randomness of numbers generated in such a fashion was checked by sorting 5000 of them into 50 categories on the

basis of the first 2 digits. A χ^2 was computed to determine the fit of the obtained distribution to a theoretical one consisting of 100 scores in each of the 50 categories. The obtained χ^2 of 47.83 is extremely close to the 49.332 value, which is the theoretical median of the χ^2 distribution with 50 degrees of freedom.

(*b*) In order to obtain the random deviates (the individual random scores from the appropriate population), the random numbers obtained in the above fashion were considered to be numbers between 0 and 1 and interpreted as the cumulative probability for a particular score from the prescribed population. From a table entered in the machine, a random deviate having that probability was selected. This is identical with the procedure one uses in entering the ordinary z table to determine the score below which, say, 97.5% of the scores in the distribution lie. The obtained value, 1.96, is the deviate corresponding to that cumulative percentage. The distribution of such deviates from a normal population obtained by using a large random sample of probabilities is normally distributed. Similar tables can be constructed for other populations. The populations selected for this study were the normal, the exponential (**J**-shaped with a skew to the right) having a density function of $y = e^{-x}$, and the rectangular or uniform distribution. These distributions represent extremes of skewness and flatness to compare with the normal. The tables of deviates corresponding to each of the selected distributions were contained internally in the computer and were so arranged that the mean of each distribution was 0 and the variance 1. To verify these values, population means and variances based on samples of 5000 deviates from each of the three populations were estimated by the usual formulas. The results were for the normal distribution a sample mean of .0024 and a variance of 1.0118, for the exponential a mean of .0128 and a variance of 1.0475, and for the rectangular a mean of $-.0115$ and a variance of .9812. All of these results could quite easily have arisen from random sampling from distributions having the assumed characteristics. To change the size of the variance of the population, all deviates were multiplied when necessary by a constant, in this case, the number 2. The resulting distribution has a mean of 0 and variance of 4. The only variances used in this study were 1 and 4.

(*c*) The sample sizes selected were 5 and 15.

(*d*) The formula used for the computation of t was the following:

$$t = \frac{M_1 - M_2}{\sqrt{\dfrac{\Sigma X_1^2 - N_1 M_1^2 + \Sigma X_2^2 - N_2 M_2^2}{N_1 + N_2 - 2}\left(\dfrac{1}{N_1} + \dfrac{1}{N_2}\right)}}$$

where M_1 and M_2 are the means of the first and second samples and N_1 and N_2 are the respective sample sizes. This expression, or an equivalent statement of it, is found in any statistics book and is undoubtedly employed in a preponderance of the research in which a t test involving nonrelated means is used. As pointed out in most statistic texts, this test is not appropriate when variances are different. Tests are available which are more or less legitimate under these conditions, but a certain amount of approximation is involved in them. It was felt, however, that

the ordinary *t* test might under some conditions be as good an approximation as the more complex forms of *t* tests and that a verification of this notion was desirable. In addition, the above formula makes use of a pooled estimate of variance for the error term and in this respect is similar to the *F* test of analysis of variance. Because of this fact, certain results can be generalized from the *t* to the *F* test.

To summarize, random samples were drawn from populations which were either normal, rectangular, or exponential with means equal to 0 and variances of 1 or 4. For several combinations of forms and variances, *t* tests of the significance of the difference between sample means were computed using combinations of the sample sizes 5 and 15. For each of these combinations, frequency distributions of sample *t*'s were obtained on the IBM 650 Electronic Computer.

Results

The results of the sampling study will be presented in part as a series of frequency distributions in the form of bar graphs of the obtained distribution of *t*'s for a particular condidition. Upon these have been superimposed the theoretical *t* distribution curve for the appropriate degrees of freedom. This furnishes a rapid comparison of the extent to which the empirical distribution conforms to the theoretical.

First we shall consider those combinations possible when both of the samples are from normal distributions but variances and sample sizes may vary. Next will be considered the results of sampling from non-normal distributions, but both samples are from the same type of distribution. Finally we deal with the results of sampling from two different kinds of populations, for example, one sample from the normal distribution, and another from the exponential.

Potentially, a very large number of such combinations are possible. Limitations of the time available on the computer necessitated a paring down to a reasonable number. Although the computer is relatively fast when optimally programmed, it nevertheless required almost an hour, on the average, to complete a frequency distribution of 1000 *t*'s. The combinations presented here are those which seemed most important at the time the study was made.

As a measure of the effect of violation of assumptions, the percentage of obtained *t*'s which exceed the theoretical values delineating the middle 95% of the *t* distribution is used. For 8, 18, and 28 *df* which arise in the present study, the corresponding values are respectively ± 2.262, ± 2.101, and ± 2.048. If the assumptions are met, and if the null hypothesis of equality of means is true, 5% of the obtained *t*'s should fall outside these limits. The difference between this nominal value and the actual value obtained by sampling should be a useful measure of the degree to which violation of assumptions changes the distribution of *t* scores. There is, of course, a random quality to the obtained percentage of *t*'s falling outside the theoretical limits. Hence, the obtained value should be looked upon as an approximation to the true value which should lie nearby.

In the figures and in the text, the various combinations of population, variance, and sample size will be represented symbolically in the following form: $E(0, 1)5\text{-}N(0, 4)15$. Here the letters *E*, *N*, and *R* refer to the population from which the sample was drawn, *E* for exponential-

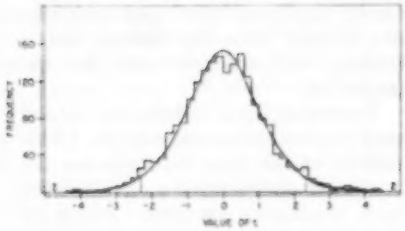

FIG. 1. Empirical distribution of t's from $N(0, 1)5$-$N(0, 1)5$ and theoretical distribution with 8 *df*.

N for normal, and R for rectangular. The first number in the parenthesis is the mean of the population distribution, in all cases zero, while the second number is the variance. The number following the parenthesis is the sample size for that particular sample. In the example above, the first sample is of Size 5 from an exponential distribution having a variance of 1. The second sample is from a normal distribution with variance of 4 and the sample size is 15.

Sampling from Normal Distributions

In order to justify the random sampling approach utilized in this study, and partly to confirm the faith placed in the tabled values of the mathematical statisticians, the initial comparisons are between the theoretical distributions and the obtained distributions with assumptions inviolate. Figures 1 and 2 exhibit the empirical distributions of t's when both samples are taken from the same normal distribution with zero mean and unit variance—designated $N(0, 1)$. In Fig. 1 both samples are of Size 5, while both are 15 in Fig. 2. The theoretical curves, one for 8 *df*, the other for 28, represent quite well the obtained distributions. Ordinates approximately two units from the mean of the theoretical distributions mark off the respective 5% limits for rejecting the null hypothesis. In Fig. 1,[1] 5.3% of the obtained t's fall outside these bounds, while in Fig. 2 only 4.0% of the sample t's are in excess. Since in both cases the expected value is exactly 5%, we must attribute the discrepancy to random sampling fluctuations. The size of these discrepancies should be useful measures in evaluating the discrepancies which will be encountered under other conditions of sampling. For examples of 2000 t's a discrepancy as large as 1% from the nominal 5% value evidently occurs frequently, and for this reason should not be considered as evidence to reject the theoretical distribution as an approximation to the empirical one.

As an initial departure from the simplest cases just presented, Fig. 3 compares theoretical and empirical distributions when samples are taken from the same $N(0, 1)$ population, but the first sample size is 5, the second is 15—that is, $N(0, 1)5$-$N(0, 1)15$. While this in no sense is a violation of the assumptions of the t test, it is interesting to note that again sampling fluctuations have produced an empirical distribution with 4.0% of the t's falling outside the nominal 5% limits.

[1] The numbers in the tails of some of the figures report the number of obtained t's falling outside the boundaries.

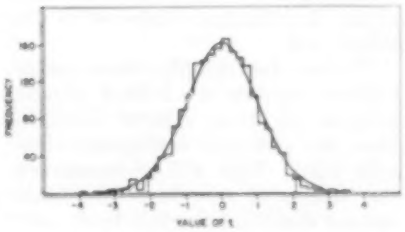

FIG. 2. Empirical distribution of t's from $N(0, 1)15$-$N(0, 1)15$ and theoretical distribution with 28 *df*.

The violation of the assumption of homogeneity of variance has effects as depicted in Fig. 4. Here the obtained distribution is based upon two samples of Size 5, one from $N(0, 1)$ and the other from $N(0, 4)$. The fit is again seen to be close between theoretical and empirical distributions, and 6.4% of the obtained *t*'s exceed the theoretical 5% limits. By increasing the sample size to 15, a distribution results (not shown here) for which only 4.9% of the *t*'s fall outside the nominal limits. It would seem that increasing the sample size produces a distribution which conforms rather closely to the *t* distribution. As will be seen later, this is a quite general result based upon mathematical considerations, the implications of which are important to the argument. For the moment it is evident that differences in variance at least in the ratio of 1 to 4 do not seriously affect the accuracy of probability statements made on the basis of the *t* test.

This last conclusion is true only so long as the size of both samples is the same. If the variances are different, with the present set of conditions there are two combinations of variance and sample size possible. In one case the first sample may be of Size 5 and drawn from the population with the smaller variance, while

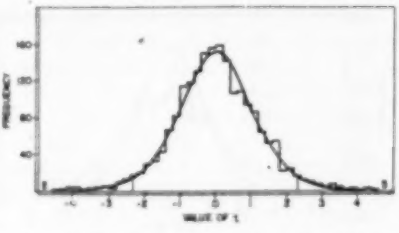

FIG. 4. Empirical distribution of *t*'s from $N(0, 1)5$-$N(0, 4)5$ and theoretical distribution with 8 *df*.

the second sample of Size 15 is drawn from the population having the larger variance—$N(0, 1)5$-$N(0, 4)15$. In the second case the small sample size is coupled with the larger variance, the larger sample size with the smaller variance—$N(0, 4)5$-$N(0, 1)15$. The respective results of such sampling are presented in Fig. 5 and 6. The empirical distributions are clearly not approximated by the *t* distribution. For the distribution of Fig. 5, only 1% of the obtained *t*'s exceed the nominal 5% values, while in Fig. 6, 16% of the *t*'s fall outside those limits.

There are good mathematical reasons why a difference in sample size should produce such decided discrepancies when the variances are unequal. Recall that $\Sigma(X-M)^2/(N-1)$ is an estimate of the variance of the

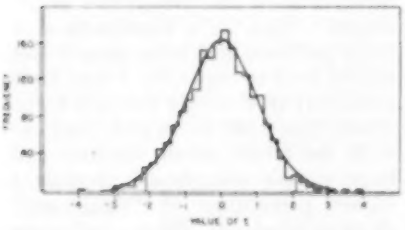

FIG. 3. Empirical distribution of *t*'s from $N(0, 1)5$-$N(0, 1)15$ and theoretical distribution with 18 *df*.

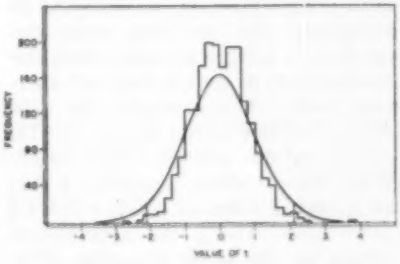

FIG. 5. Empirical distribution of *t*'s from $N(0, 1)5$-$N(0, 4)15$ and theoretical distribution with 18 *df*.

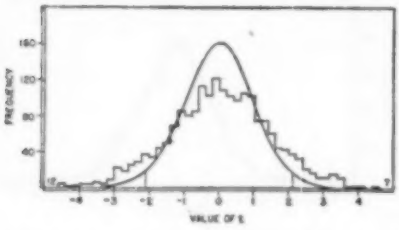

FIG. 6. Empirical distribution of t's from $(N(0, 4)5\text{-}N(0, 1)15$ and theoretical distribution with 18 df.

population from which the sample is drawn. Hence, $\Sigma(X-M)^2$ will in the long run be equal to $(N-1)\sigma^2$. The formula used in this study for computing t makes use of this fact and, in addition, under the assumption that the variances of the populations from which the two samples are drawn are equal, pools the sum of the squared deviations from the respective sample means to get a better estimate. That is $\Sigma(X_1-M_1)^2+\Sigma(X_2-M_2)^2$ is an estimate of $(N_1-1)\sigma_1^2+(N_2-1)\sigma_2^2$. If $\sigma_1^2=\sigma_2^2=\sigma^2$ (homogeneity of variance), then the sums estimate $(N_1+N_2-2)\sigma^2$. Hence,

$$\frac{\Sigma(X_1-M_1)^2+\Sigma(X_2-M_2)^2}{N_1+N_2-2} \quad [1]$$

is an estimate of σ^2. If $\sigma_1^2\neq\sigma_2^2$ the estimating procedure is patently illegitimate, the resulting value depending in a large measure upon the combination of sample size and variance used. For example, the case $N(0, 1)5\text{-}N(0, 4)15$ has $N_1=5$, $N_2=15$, $\sigma_1^2=1$, $\sigma_2^2=4$, and $N_1+N_2-2=18$. With these values, Formula 1 has an expected value of $[(4\cdot1)+(14\cdot4)]/18=3.33$. Using the appropriate values for the other situation, $N(0, 4)5\text{-}N(0, 1)15$, the result of formula 1 is $[(4\cdot4)+(14\cdot1)]/18=1.67$. This means that on the average, the denominator for the t test will be larger for the first case than for the second. If the sample differences between means were of the same magnitude for the two cases, obviously more "significant" t's would emerge when the denominator is smaller. It so happens that when this latter condition exists, the variance of the numerator also tends to be greater than in the other condition, a fact which accentuates the differences between the two empirical distributions.

Welch (1937) has shown mathematically that in the case of sample sizes of 5 and 15, a state which prevails here, the percentage of t's exceeding the nominal 5% value varies as a function of the ratio of the two population variances and can be as low as 0% and as high as 31.3%. If $N_1=N_2$ there is never much bias, except perhaps in the case in which the sample sizes are both 2. For $N_1=N_2=10$, the expected value of the percentage of t's exceeding the nominal 5% limits varies between 5% and 6.5% regardless of the difference between the variances. For larger sample sizes, the discrepancy tends to be even less.

Since the pooling procedure for estimating the population variance is used in ordinary analysis of variance techniques, it would seem that the combination of unequal variances and unequal sample sizes might play havoc with F test probability statements. That is, a combination of large variance and large sample size should tend to make the F test more conservative than the nominal value would lead one to expect, and, as with the t test, small variance and large sample size should produce a higher percentage of "significant" Fs than expected. These conclusions are based upon a very simple extension to more than two samples of the explanation for the behavior of the

t test probabilities with unequal sample sizes.

A more sophisticated mathematical handling of the problem by Box (1954a) reaches much the same conclusions for the simple-randomized analysis of variance. In a table in his article are given exact (i.e., mathematically determined) probabilities of exceeding the 5% point when variances are unequal. In this case, sampling is assumed to be from normal distributions. If the sample sizes are the same, the probability given for equal sample sizes range from 5.55% to 7.42%, for several combinations of variances, and numbers of samples. If, when variances are different, the samples are of different sizes, large discrepancies from the nominal values result. Combining large sample and large variance lessens the probability of obtaining a "significant" result to much less than 5%, just as we have seen for the *t* test. In a subsequent article, Box (1954b) presents some results from two-way analysis of variance. Since these designs generally have equal cell frequencies the results are not too far from expected. His figures all run within 2% of the 5% value expected if all assumptions were met.

It would seem then that both empirically and mathematically there can be demonstrated only a minor effect on the validity of probability statements caused by heterogeneity of variance, provided the sizes of the samples are the same. This applies to the *F* as well as the *t* test. If however, the sample sizes are different, major errors in interpretation may result if normal curve thinking is used.

Sampling from Identical Non-Normal Distributions: (Equal Variances)

Let us now proceed to violate the other main assumption, that of normality of distribution from which sampling takes place. At this time we will consider the *t* distributions arising when both samples are taken from the same non-normal distribution. The distributions shown here, and all subsequent ones, are based upon only 1000 *t*'s, and hence will exhibit somewhat more column to column fluctuation than the preceding distributions.

Figure 7 compares the theoretical *t* distribution and the empirical distribution obtained from two samples of Size 5 from the exponential distribution—*E*(0, 1)5-*E*(0, 1)5. The fit is fairly close, but the proportion of cases in the tails seems less for the empirical distribution than for the theoretical. By count, 3.1% of the obtained *t*'s exceed the nominal 5% values—that is, the test in this case seems slightly conservative. If both sample sizes are raised to 15 (distribution not shown here), the corresponding percentage of obtained *t*'s is 4.0%. While this is probably not an appreciably better fit than for samples of Size 5, we shall see later that there are theoretical reasons to suspect that increasing the sample size should better the approximation of the empirical curve by the theoretical no matter what the parent population may be.

If both samples are of Size 5 from the same rectangular distribution—*R*(0, 1)5-*R*(0, 1)5—the result is as

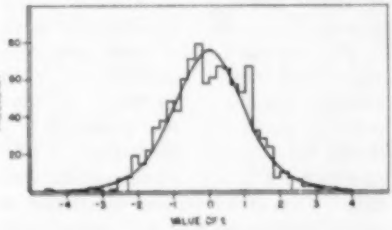

FIG. 7. Empirical distribution of *t*'s from *E*(0, 1)5-*E*(0, 1)5 and theoretical distribution with 8 *df*.

depicted in Fig. 8. The fit of theoretical curve to empirical data here is as good as any thus far observed. The percentage of obtained t's exceeding the 5% values is 5.1% in this particular case. For the case in which the sample sizes are both 15 (not shown here), the fit is equally good, with 5.0% of the cases falling outside of the nominal 5% bounds.

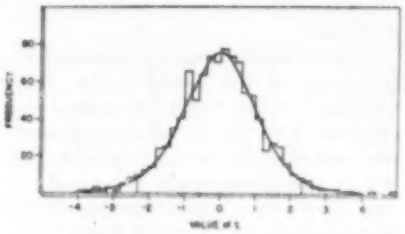

FIG. 8. Empirical distribution of t's from $R(0, 1)5$-$R(0, 1)5$ and theoretical distribution with 8 *df*.

Sampling from Non-Normal Distributions: (Unequal Variances)

We may assume that if the variances are unequal, and at the same time the sample sizes are different, the resulting distributions from non-normal populations will be affected in the same way as the distributions derived from normal populations, and for the same reasons. These cases will not be considered.

If sampling is in sizes of 5 from two exponential distributions, one with a variance of 1, and the other of 4, a skewed distribution of obtained t's emerges (not shown here). We shall discover that a skewed distribution of t's generally arises when the sampling is from distributions which are different in degree of skewness or asymmetry. (For an explanation, see discussion of $E(0, 1)5$-$N(0, 1)5$ below.) Apparently, the effect of increasing the variance of the exponential distribution as in the present case—$E(0, 1)5$-$E(0, 4)5$—is to make the negative sample means arising from the distribution with larger variance even more negative than those from the distribution with smaller variance. In terms of percentage exceeding the nominal 5% limits for this case, the value is 8.3%, of which 7.6% comes from the skewed tail. This combination of variances and distribution was not tested with larger samples, but we shall see when comparing exponential and normal distributions that an increase in the sample size decreases the skew of the obtained t distribution there. Theoretically, this decrease should occur in almost all cases, including the present one.

The result is much less complicated if, while variances are different, the sampling is from symmetrical rectangular distributions—$R(0, 1)$-$R(0, 4)5$. For this small sample situation, (not illustrated), there occurs a distribution of obtained t's having 7.1% of the values exceeding the nominal 5% points. This is roughly the same magnitude as the corresponding discrepancy from normal distributions. For the normal, it will be recalled that an increase of the sample sizes to 15 decreased the obtained percentage to 4.9%. There is no reason to believe that increasing the size of the rectangular samples would not have the same effect. However, time did not permit the determinaiton of this distribution.

Sampling from Two Different Distributions

By drawing the first sample from a distribution having one shape, and by drawing the second from a distribution having another shape (other than shape differences arising from heterogeneity of variance), yet another way has been found to do violence to the integrity of the as-

sumptions underlying the *t* test. Perhaps the least violent of these happenings is that in which at least one of the populations is normal.

When one sample is from the exponential distribution and the other from the normal, the interesting result shown in Fig. 9 occurs. This is the small sample case—$E(0, 1)5\text{-}N(0,1)5$. It will be recalled that for skewed distributions the mean and median are at different points. In the exponential distributions, for example, the mean is at the 63rd centile. If samples from the exponential distribution are small, there will be a tendency for the sample mean to be less than the population mean, obviously since nearly two thirds of the scores are below that mean. Since the population mean of the present distributions is 0, the result will be a preponderance of negative sample means for small samples. If the other sample is taken from a symmetrical distribution, which would tend to produce as many positive as negative sample means, the resulting distribution of obtained *t*'s would not balance about its zero point, an imbalance exacerbated by small samples. In Fig. 9, 7.1% of the obtained cases fall outside the 5% limits, with most, 5.6%, lying in the skewed tail. The effect of increasing the sample size to 15 is to normalize the distribution considerably; the resulting curve, Fig. 10, is fairly well approximated by the *t* distribution. One of the tails, however, does contain a disproportionate share of the cases, 4.2% to 0.9% for the other tail, or a total of 5.1% falling outside the nominal 5% limits. Nevertheless, the degree to which the theoretical and empirical distributions coincide under these conditions is striking. It seems likely that if both samples were each of Size 25, the resulting sample distribution of *t*'s would be virtually indistinguishable from the *t* distribution for 48 *df*, or the next best thing, the normal curve itself. To test this hypothesis, an additional empirical *t* distribution based on sample sizes of 25 from these same exponential and normal populations was obtained (not shown here). The results nicely confirm the presumption. Comparison with the usual 5% values reveals 4.6% of the empirical *t*'s surpassing them. Whereas with the smaller samples the ratio of *t*'s in the skewed tail to those in the other tail is roughly 80:20, the corresponding ratio for the larger sample case is 59:41. Clearly, the increase in sample sizes has tended to normalize the distribution of *t*'s.

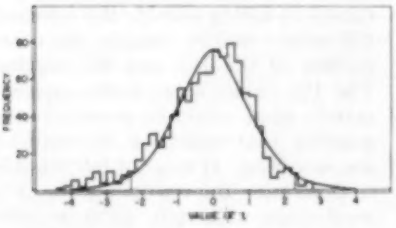

FIG. 9. Empirical distribution of *t*'s from $E(0, 1)5\text{-}N(0, 1)5$ and theoretical distribution with 8 *df*.

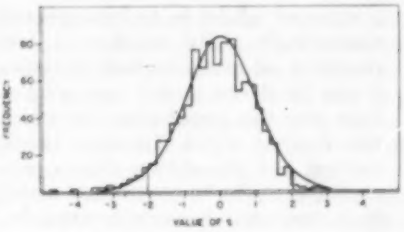

FIG. 10. Empirical distribution of *t*'s from $E(0, 1)15\text{-}N(0, 1)15$ and theoretical distribution with 28 *df*.

For these conditions, involving rather drastic violation of the mathematical assumptions of the test, the *t* test has been observed to fare well with an adequate sample size. Such

a state of affairs is to be expected theoretically. By invoking a few theorems of mathematical statistics it can be shown that if one samples from any two populations for which the Central Limit Theorem holds, (almost any population that a psychologist might be confronted with), no matter what the variances may be, the use of equal sample sizes insures that the resulting distribution of t's will approach normality as a limit. It would appear from the present results that the approach to normality is rather rapid, since samples of sizes of 15 are generally sufficient to undo most of the damage inflicted by violation of assumptions. Only in extreme cases, such as the last which involves distributions differing in skew, would it seem that slightly larger sizes are prescribed. Thus it would appear that the t test is functionally a distribution-free test, providing the sample sizes are sufficiently large (say, 30, for extreme violations) and equal.

The distributions arising when sampling is from the normal and the rectangular distributions—$N(0, 1)5$-$R(0, 1)5$ and $N(0, 1)15$-$R(0, 1)15$—would further tend to substantiate this claim. The respective percentages exceeding the 5% nominal values are 5.6% and 4.6% from the empirical distributions for these cases, the distributions of t's being symmetrical and close to the theoretical (not shown).

The only other combination examined in the sampling study is the uninteresting case of exponential and rectangular distributions. This distribution (not shown) is again skewed with the effect of increase of sample size from 5 to 15 to cut down the skew and to decrease the percentage of cases falling outside the theoretical 5% values from 6.4% to 5.6%. For those cases falling outside the nominal 5% values, the ratio is 79:21 for the smaller samples. This is changed to 69:31 for the sample size of 15. Here again it would seem that larger sample sizes would be required to insure the validity of probability statements utilizing the t distribution as a model.

The results of the total study are summarized in Table 1 which gives for each combination of population, variance, and sample size (a) the percentage of obtained t's falling outside the nominal 5% probability limits of the ordinary t distribution, and (b) the percentage of obtained t's falling outside the 1% limits. The combinations are represented symbolically as before. The table is divided into two parts, the first part presenting information on the empirical distributions which are intrinsically symmetrical. The second part is based upon the intrinsically nonsymmetrical distributions, additional information in this section of the table being the percentage of obtained t's falling in the larger of the tails. The percentage for the smaller tail may be obtained by subtraction of the percentage in the larger tail from the total.

Certain implications of the table should be discussed. In the Norton study, more severe distortions sometimes occurred with significance levels of 1% and .1% than appeared with the 5% level. The inclusion in Table 1 of the percentages of obtained t's falling outside the nominal 1% values makes possible the comparison of the 1% and 5% results. The 1% values seem to be approximately what would be expected considering that sampling fluctuations are occurring. It was not felt feasible to determine the results for the .1% level since with only 1000 or 2000 cases the number of obtained t's falling outside the prescribed limits was negligible in most cases. It is pos-

sible, however, that the distortions in the apparent level of significance are more drastic for the smaller α values.

All the results and discussion have been limited thus far to the two-tailed t test. With notable exceptions, the conclusions we have reached can be applied directly to the one-tailed t test as well. The exceptions involve those distributions which are intrinsically *asymmetric* (see Table 1). In these distributions a preponderance of the obtained t's fall in one tail. Depending upon the particular tail involved in the one-tailed test the use of t should produce too many or too few significant results when sampling is from a combination of populations from which an asymmetric t distribution is expected. It seems impossible to make any simple statements about the behavior of the tails in the general case of asymmetric t distribution except to say that such distributions are expected whenever the skew of the two parent populations is different. The experimenter must determine for each particular instance the direction of skew of the expected distribution and act accordingly. Table 1 gives for the intrinsically asymmetric distributions the total percentage of obtained t's falling outside the theoretical 5% and 1% limits and the percentage in the larger tail. From these values can be assessed the approximate magnitude of the bias incurred when a one-tailed test is used in specific situations.

TABLE 1

OBTAINED PERCENTAGES OF CASES FALLING OUTSIDE THE APPROPRIATE TABLED t VALUES FOR THE 5% AND 1% LEVEL OF SIGNIFICANCE

Symmetric Distributions	Obtained Percentage at 5% level	Obtained Percentage at 1% level
$N(0, 1)5$-$N(0, 1)5$	5.3	0.9
$N(0, 1)15$-$N(0, 1)15$	4.0	0.8
$N(0, 1)5$-$N(0, 1)15$	4.0	0.6
$N(0, 1)5$-$N(0, 4)5$	6.4	1.8
$N(0, 1)15$-N$(0, 4)15$	4.9	1.1
$N(0, 1)5$-$N(0, 4)15$	1.0	0.1
$N(0, 4)5$-$N(0, 1)15$	16.0	6.0
$E(0, 1)5$-$E(0, 1)5$	3.1	0.3
$E(0, 1)15$-$E(0, 1)15$	4.0	0.4
$R(0, 1)5$-$R(0, 1)5$	5.1	1.0
$R(0, 1)15$-$R(0, 1)15$	5.0	1.5
$R(0, 1)5$-R$(0, 4)5$	7.1	1.9
$N(0, 1)5$-$R(0, 1)5$	5.6	1.0
$N(0, 1)15$-$R(0, 1)15$	5.6	1.1

Asymmetric Distributions	Obtained Percentage at 5% level		Obtained Percentage at 1% level	
	Total	Larger Tail	Total	Larger Tail
$E(0, 1)5$-$N(0, 1)5$	7.1	5.6	1.9	1.9
$E(0, 1)15$-$N(0, 1)15$	5.1	4.2	1.4	1.2
$E(0, 1)25$-$N(0, 1)25$	4.6	2.7	1.3	1.1
$E(0, 1)5$-$R(0, 1)5$	6.4	5.0	3.3	2.5
$E(0, 1)15$-$R(0, 1)15$	5.6	3.9	1.6	1.2
$E(0, 1)5$-$E(0, 4)5$	8.3	7.6	1.7	1.7

DISCUSSION AND CONCLUSIONS

Having violated a number of assumptions underlying the t test, and finding that, by and large, such violations produce a minimal effect on the distribution of t's, we must conclude that the t test is a remarkably *robust* test in the technical sense of the word. This term was introduced by Box (1953) to characterize statistical tests which are only inconsequentially affected by a violation of the underlying assumptions. Every statistical test is in part a test of the assumptions upon which it is based. For example, the null hypothesis of a particular test may be concerned with sample means. If, however, the assumptions underlying the test are not met, the result may be "significant" even though the population means are the same. If the statistical test is relatively insensitive to violations of the assumptions other than the null hypothesis, and, hence, if probability statements refer pri-

marily to the null hypotheses, it is said to be robust. The t and F tests apparently possess this quality to a high degree.

In this particular context, an important example of a test lacking robustness is Bartlett's test for homogeneity of variance (Bartlett, 1937). Box (1953) has shown that this test is extremely sensitive to non-normality and will under some conditions be prone to yield "significant" results even if variances are equal. For example, Box tables a number of exact probabilities of exceeding the 5% normal theory significance level in the Bartlett test for various levels of λ_2, the kurtosis parameter, for different quantities of variances being compared. As an extreme case, if $\lambda_2 = 2$ (i.e., a peaked distribution) with 30 variances being tested, the probability of rejecting the hypothesis at the nominal .05 level is actually .849. If $\lambda_2 = -1$ (i.e., a flat distribution), the probability is .00001. Note that in both these cases, all variances are actually equal. Box, realizing that in the case of equal sample sizes the analysis of variance is affected surprisingly little by heterogeneous variance and nonnormality, concludes that the use of the nonrobust Bartlett test to "make the preliminary test on variances is rather like putting out to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!" Apparently, as reported in this same article, other commonly used tests for evaluating homogeneity are subject to the same weakness.

We may conclude that for a large number of different situations confronting the researcher, the use of the ordinary t test and its associated table will result in probability statements which are accurate to a high degree, even though the assumptions of homogeneity of variance and normality of the underlying distributions are untenable. This large number of situations has the following general characteristics: (*a*) the two sample sizes are equal or nearly so, (*b*) the assumed underlying population distributions are of the same shape or nearly so. (If the distributions are skewed they should have nearly the same variance.) If these conditions are met, then no matter what the variance differences may be, samples of as small as five will produce results for which the true probability of rejecting the null hypothesis at the .05 level will more than likely be within .03 of that level. If the sample size is as large as 15, the true probabilities are quite likely within .01 of the nominal value. That is to say, the percentage of times the null hypothesis will be rejected when it is actually true will tend to be between 4% and 6% when the nominal value is 5%.

If the sample sizes are unequal, one is in no difficulty provided the variances are compensatingly equal. A combination of unequal sample sizes and unequal variances, however, automatically produces inaccurate probability statements which can be quite different from the nominal values. One must in this case resort to different testing procedures, such as those by Cochran and Cox (1950), Satterthwaite (1946), and Welch (1947). The Welch procedure is interesting since it has been extended by Welch (1951) to cover the simple randomized analysis of variance which suffers the same defect as the t test when confronted with both unequal variance and unequal sample sizes. The Fisher-Behrens procedure suggested by many psychologically oriented statistical textbooks has had its validity questioned (Bartlett, 1936) and, hence, is ignored by

some statisticians (e.g., Anderson & Bancroft, 1952, p. 82).

If the two underlying populations are not the same shape, there seems to be little difficulty if the distributions are both symmetrical. If they differ in skew, however, the distribution of obtained t's has a tendency itself to be skewed, having a greater percentage of obtained t's falling outside of one limit than the other. This may tend to bias probability statements. Increasing the sample size has the effect of removing the skew, and, due to the Central Limit Theorem and others, the normal distribution is approached by this maneuver. By the time the sample sizes reach 25 or 30, the approach should be close enough that one can, in effect, ignore the effects of violations of assumptions except for extremes. Since this is so, the t test is seen to be functionally nonparametric or distribution-free. It also retains its power in some situations (David & Johnson, 1951). There is, unfortunately, no guarantee that the t and F tests are uniformly most powerful tests. It is possible, even probable, that certain of the distribution-free methods are more powerful than the t and F tests when sampling is from some unspecified distributions or combination of distributions. At present, little can be said to clarify the situation. Much more research in this area needs to be done.

Since the t and F tests of analysis of variance are intimately related, it can be shown that many of the statements referring to the t test can be generalized quite readily to the F test. In particular, the necessity for equal sample sizes, if variances are unequal, is important for the same reasons in the F test of analysis of variance as in the t test. A number of the cited articles have demonstrated both mathematically and by means of sampling studies that most of the statements we have made do apply to the F test. It is suggested that psychological researchers feel free to utilize these powerful techniques where applicable in a wider variety of situations, the present emphasis on the nonparametric methods notwithstanding.

REFERENCES

Anderson, R. L., & Bancroft, T. A. *Statistical theory in research.* New York: McGraw-Hill, 1952.

Bartlett, M. S. The effect of non-normality on the t-distribution. *Proc. Camb. Phil. Soc.*, 1935, **31**, 223–231.

Bartlett, M. S. The information available in small samples. *Proc. Camb. Phil. Soc.*, 1936, **32**, 560–566.

Bartlett, M. S. Properties of sufficiency and statistical tests. *Proc. Roy. Soc.* (*London*), 1937, **160**, 268–282.

Box, G. E. P. Non-normality and tests on variances. *Biometrika*, 1953, **40**, 318–335.

Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Ann. of math. Statist.*, 1954, **25**, 290–302. (a)

Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effects of inequality of variance and of correlation between errors in the two-way classification. *Ann. of math. Statist.*, 1954, **25**, 484–498. (b)

Box, G. E. P., & Andersen, S. L. Permutation theory in the derivation of robust criteria and the study of departures from assumption. *J. Roy. Statist. Soc.* (*Series B*), 1955, **17**, 1–34.

Cochran, W. G., & Cox, G. M. *Experimental designs.* New York: Wiley, 1950.

Daniels, H. E. The effect of departures from ideal conditions other than non-normality on the t and z tests of significance. *Proc. Camb. Phil. Soc.*, 1938, **34**, 321–328.

David, F. N., & Johnson, N. L. The effect of non-normality on the power function of the F-test in the analysis of variance. *Biometrika*, 1951, **38**, 43–57.

Gayen, A. K. The distribution of the vari-

ance ratio in random samples of any size drawn from non-normal universes. *Biometrika*, 1950, **37**, 236–255. (a)

GAYEN, A. K. Significance of difference between the means of two non-normal samples. *Biometrika*, 1950, **37**, 399–408. (b)

HORSNELL, G. The effect of unequal group variances on the *F*-test for the homogeneity of group means. *Biometrika*, 1953, **40**, 128–136.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.

NORTON, D. W. An empirical investigation of some effects of non-normality and heterogeneity on the *F*-distribution. Unpublished doctoral dissertation, State Univer. of Iowa, 1952.

PEARSON, E. S. The analysis of variance in the case of non-normal variation. *Biometrika*, 1931, **23**, 114–133.

QUENSEL, C. E. The validity of the *Z*-criterion when the variates are taken from different normal populations. *Skand. Aktuarietids*, 1947, **30**, 44–55.

SATTERTHWAITE, F. E. An approximate distribution of estimates of variance components. *Biomet. Bull.*, 1946, **2**, 110–114.

WELCH, B. L. The significance of the difference between two means when the population variances are unequal. *Biometrika*, 1937, **29**, 350–362.

WELCH, B. L. The generalization of Student's problem when several different population variances are involved. *Biometrika*, 1947, **34**, 28–35.

WELCH, B. L. On the comparison of several mean values: An alternative approach. *Biometrika*, 1951, **38**, 330–336.

(January 28, 1959)

PSYCHOLOGICAL BULLETIN
VOL. 57, NO. 1, 1960

THE INFLUENCE OF SITUATIONAL AND INTERPERSONAL VARIABLES IN PROJECTIVE TESTING

JOSEPH MASLING
Syracuse University

Since the advent of projective techniques, psychologists have devoted considerable energy to the creation and utilization of various methods for assessing the "deeper layers" of personality. The impressive results often obtained with these instruments have led to a proliferation of projective methods, most of which are greeted with enthusiasm and validated by endorsement. The early, uncritical acceptance of projective testing produced studies which were mainly concerned with the characteristics of such clinical populations as delinquents, psychotics, and organics. In general, the assumption was made that a projective test was as single-minded as the X ray (Frank, 1939) revealing information only about the patient without in any way being influenced by the person who administered the test, the method of administration, or the situation in which it was used. In the words of Cronbach (1956), "Test research has been dominated by the Galtonian view that the test is a sample of the subject's responses to a standardized nonpersonal stimulus" (p. 175).

However, with growing sophistication regarding the nature of psychological testing, a number of writers have explicitly commented on the influence in projective testing of factors other than the *S*'s personality. One of the first to delineate the subjective factors in Rorschach testing was Schactel (1945) who described four common elements in the Rorschach situation: the relationship of the *E* and *S*; the assignment of the task by the *E* to the *S*; the *E*'s need to interpret the *S*'s behavior; and the specific qualities of the task, such as the ambiguity or the lack of familiarity with the stimuli. Miller (1953), Sarason (1954), and Luchins (1947) have also indicated the subtle ways in which subjective forces may influence the course of a projective testing situation.

It is the purpose of this paper to review the considerable evidence regarding situational and interpersonal influences in projective testing. Experiments dealing with the effects of these factors on interviewing or intelligence testing and those studies concerning the effects of psychotherapy, psychosurgery, and individual vs. group tests will not be reviewed, since these issues are considered to be part of essentially different problems. Four arbitrary, overlapping categories will be used to present these studies: method of administration, the testing situation, examiner influence, and subject influence.

THE INFLUENCE OF METHOD OF ADMINISTRATION

Instructions to make a good or bad impression. The ability of projective tests to withstand attempts by *S*s to disguise or alter their "real" responses has been investigated several times. The usual procedure in these studies is to test the same *S*s several times under varying instructions; comparisons are then made between the test responses produced under standard instructions with those yielded by experimental instructions.

Fosberg (1938, 1941, 1943) has reported on the process of trying to

produce a good or bad impression on the Rorschach. In one study (Fosberg, 1938) a husband and wife were each administered the Rorschach under four sets of instructions. Despite the instructions to create a given impression, it was concluded that the *Ss* were unable to avoid revealing basic aspects of their personality, the psychograms on the four examinations remaining essentially the same. In a later study (Fosberg, 1941), the Rorschach was given four times under different instructions to 25 male and 25 female *Ss*. A special experimental group of 16 *Ss* took the fourth examination under instructions to look for particular determinants. When Fosberg compared group means produced under the different instructions he found few consistent differences. There was little change in the test as a whole, only the content of the responses showing marked changes. One reason for the failure of these instructions to produce differences in responses is that each *S* defined for himself the manner in which to deceive the *E*, so that six *Ss* increased their responses in order to make a good impression while four increased their responses in order to make a poor impression. While there was no consistency among all *Ss* on how to create an artificial image for the *E*, most *Ss* felt that they could falsify their reactions by adopting a particular set. Of special importance to this review is the fact that in trying to make a bad impression several *Ss* seemed to concentrate on the *E*, rather than the test; three *Ss* annoyed the *E*, two acted stubborn and ornery, two proceeded very slowly, and four paid little attention (Fosberg, 1943).

Fosberg's studies were essentially repeated by Carp and Shavzin (1950) who also found that taking the Rorschach under instructions to make a good impression ("you are in a state hospital and the results on this test may help get you out") and under instructions to make a bad impression ("you are to be drafted for the Army and the results on this test may help keep you out") produce no significant group differences (except for the *z* score, significant at the .05 level). "This does not mean, however, that no changes were produced. The data clearly showed the differences. But the direction taken was so diverse, among the individual *Ss*, that they were balanced out in the analysis" (p. 232). The authors directly challenged Fosberg's conclusions that the Rorschach could not be manipulated by the *Ss*. "On the contrary, this study shows that there are some subjects who *can* manipulate their responses, who *can* vary their personality picture as reflected by the Rorschach, under instructions to make 'good' or 'bad' impressions" (p. 233).

Weisskopf and Dieppa (1951) administered three cards of the TAT to hospitalized psychoneurotic veterans, giving the standard instructions in one administration, asking the *Ss* to give the best possible impression on another administration and the worst possible impression on the third administration. Of the nine dimensions rated by judges, five showed significant differences as a function of the instructions. When the *Ss* tried to give their worst impression, they were rated as less well-adjusted, more hostile, less willing to conform, and more spontaneous. Wallon and Webb (1957) asked naval aviation cadets to take the Rosenzweig P-F test and a sentence completion test under several variations of instructions and test structure. One group took the P-F test in a multiple choice form, another took it in standard form,

while a third group was told in taking the test to try to make the best impression. It was concluded that as the test became less ambiguous, the results more closely resembled responses produced under instructions to fake.

Instructions emphasizing particular determinants or locations. In the process of studying suggestibility, Coffin (1941) demonstrated how *S*'s set may influence his responses to the Rorschach test. The *S*s were first asked to read a fictitious article by a "Harvard professor" describing how professional men usually saw whole responses, while business men saw animals, skilled laborers saw inanimate objects and WPA employees saw details. A second group of *S*s read the article that now described professional men as seeing details, business men inanimate objects, etc. Following the reading of this article each *S* was administered six Rorschach cards. The results clearly showed the influence of the suggestion on the responses, each group tending to respond in the same direction as the socially acceptable norms. "Apparently the suggestion sets up a determining tendency operating upon the observer's perceptual and imaginal processes. This acted to direct the 'search' " (p. 62).

In a better controlled study Abramson (1951) equated two groups of college students on the basis of *W*, *D*, and *Dd* responses to the first administration of the Rorschach. One group of *S*s was then told that successful business and professional men saw whole responses, while the second group was instructed that these men saw detail responses. As a consequence of the difference in instructions, the two groups differed significantly on the second Rorschach test not only in the number of whole and detail responses but on several other determinants (*F%*, *FM*, *m*, *Hd*, *Ad*) as well, although there were no significant differences between groups on the first test. Evidently establishing a set for area will also affect those determinants dependent upon the area of the blot. The evidence of Keyes (1954) also supported the notion that the number of whole responses can be influenced by special instructions.

Hutt, Gibby, Milton, and Pottharst (1950) and Gibby (1951) investigated the effects of instructing *S*s to pay particular attention to specific aspects of the Rorschach blots. In each study various parts of the blots were emphasized to the *S*s after the standard administration of the test, but prior to the experimental administrations. A study of test-retest reliability of the Rorschach under these conditions showed certain determinants to be more stable and less resistant to change than others. "What appears to be crucial is how the individual perceives the total test situation. If we do not know this, we are likely to make serious errors in interpretation" (p. 185).

After first administering the Rorschach under standard instructions, Fabrikant (1954), in his instructions to the experimental group, stressed movement, color, shading, and texture responses. In the experimental group, 15 *S*s showed differences between first and second administrations in at least 3 of the 4 response categories, while in the control group only 3 *S*s showed significant changes in at least 3 categories.

Instructions regarding the purpose of the testing. The most impressive evidence regarding the effects of telling the *S* the purpose of the testing comes from the study of Henry and Rotter (1956), who simply told their experimental group of female

undergraduates what most college students have already presumably learned about the Rorschach test from television, the movies, and *Life* magazine: "This is a test to discover serious emotional disturbances." It was found that the experimental group gave fewer responses (at the .01 level), more good form responses (.05 level), more popular responses (.05 level) and more animal responses (.05 level) than a control group given a standard administration. It was evident that making explicit the purpose of the test produced more constriction and more attempts to be safe than leaving the purpose unstated.

A study by Calden and Cohen (1953) investigated the influence of both ego involvement and instructions regarding the nature of the Rorschach test. Half their senior high school *Ss* were given ego-involving instructions and half were given neutral instructions; one-third of the *Ss* were told that the Rorschach tested intelligence, another third that it tested "nervousness," and the last group that it measured imagination. An analysis of variance computed for 27 selected variables showed 19 differences significant at the .05 level. In general, the resulting personality pattern that emerged from the intelligence test instructions resembled the same constricted, safe picture found by Henry and Rotter, form and animal responses increasing, movement responses decreasing. "Needless to say, predictions based on 'blind' interpretations of the Rorschach protocol, without knowledge of the testing situation or the *S*'s reactions to the testing, are so much more fallible when viewed in the light of the results of this study" (p. 308–9).

The TAT has also been used to study the effects of informing *Ss* of the purpose of the test (Summerwell, Campbell, & Sarason, 1958). Four different groups were used, each group receiving one of the following instructions regarding the purpose of the test: (*a*) the usual Murray instructions, (*b*) intelligence test instructions, (*c*) projective test instructions, (*d*) neutral instructions. It was found that the neutral instructions produced significantly different emotionally toned stories than any of the other three groups, with story outcomes significantly different from the standard Murray instructions and the projective test instructions.

The manipulation of time of response. In giving the group Rorschach to 10 *Ss* under conditions of long time exposure (3 min.) and short time exposure (10 sec.), Weisskopf (1942) noted that the same personality pattern emerged under both conditions of administration. Unfortunately, adequate experimental controls were not employed. A more rigorous procedure was utilized by Siipola and Taylor (1952) who gave the Rorschach test to experimental *Ss* seated in front of a noisy automatic timer which recorded reaction time of the first response. It was concluded that the pressure induced by the timer resulted in normal *Ss* behaving under stress much as disturbed patients do in a nonstress situation. "The moral of all this seems to be that the projective process, like any other psychological process, is not immune to the influences of the specific conditions under which it operates" (p. 46).

The Influence of the Testing Situation

The designs for the studies investigating the effects of varying testing conditions take several forms. The

most rigorous of these utilizes a control group that has been given two administrations of the test to contrast with the experimental group which had experienced the special conditions between the first and second testing. If the projective test permits, some investigators prefer to counterbalance the order of presentation of the particular cards used, necessitating at least two experimental groups and two control groups. Another frequent design does not utilize a control group, the only comparison made being that of the first administration of the test with the second administration, with all differences between administrations assumed to be a function of the intervening conditions. A third procedure consists of administering a single test to groups known to differ on a particular dimension; all differences in test results are then attributed to the central, identifiable difference between the groups.

Stress. The most careful, systematic effort to induce stress was that of Lindzey (1950a, 1950b), who frustrated his experimental *Ss* by subjecting them to 10–12 hours of food deprivation, inducing them to drink a large quantity of water and then preventing them from urinating for approximately three hours, taking a blood sample in a painful way with a spring lancet, and by forcing them to fail in a group situation. As a consequence of these conditions the Rosenzweig Picture Frustration Test showed a significant increase in extra-punitive responses (Lindzey, 1950b). Of 12 predictions regarding changes in the TAT, 11 were in the expected direction, with 5 hypotheses confirmed at the .05 level of confidence or better (Lindzey & Herman, 1955). The effects of stress on the Rosenzweig P-F test were also studied by French (1950) who gave students in a social psychology class erroneous grades on an examination. Half the students who earned an A or B were given C or D, while half those earning a C or D were given an A or B. On the P-F test given immediately after the grades were returned the good students given the poor grades (the stress group) did not differ from the good students who were assigned their correct grades. However, the poor students given the erroneously high grades showed fewer intropunitive ego-defensive responses than the poor students given their correct grade.

Eichler (1951) used an elaborate device resembling an electric chair to seat his *Ss* while taking the Rorschach. They were made to wear a helmet which looked as if it could conduct electricity and were told that while taking the test they would be given shock, "the longer the time interval that elapses without the receipt of shock the more intense the next shock will be." On the basis of an administration of the Behn-Rorschach test, the experimental group was matched on five variables with a control group that took the second Rorschach under standard conditions. Judges who made a blind global rating of the Rorschach protocols found a significant difference in anxiety between the two groups. On 15 anxiety indicators, however, they found that only 4 reflected a significant difference between groups while 3 additional variables did not reach statistical significance but were in the predicted direction.

Less dramatic forms of frustration seem also to be effective in demonstrating how projective devices may reflect pretest conditions. Crandall's (1951) experimental *Ss* took tests of physical skills between administrations of the TAT and were informed that they had not meet the "norm."

A control group rested between test administrations. As a result of the failure situation the experimental *Ss*' expectation for punishment as revealed in the TAT increased significantly.

Schwartz and Kates (1957) imposed stress by giving their *Ss* a fictitious "interpretation" of their Rorschach or Behn-Rorschach responses, showing them to be poorly adjusted. When the second test was administered only 2 (*FC* and *RT*) of 16 preselected Rorschach variables showed significant differences as a function of the stress.

The best adjusted and most poorly adjusted members of NYA camp were made to fail in performing with the Rotter Level of Aspiration Board. The pre- and postfrustration TAT records of the two groups differed in several ways, the poorly adjusted group showing a decrease in themes of superiority, aggression, and emotional states. (Rodnick & Klebanoff, 1942).

Three studies did not impose stress experimentally, but utilized *Ss* in "natural" stress conditions. Klatskin (1952) gave the Rorschach test to one group of patients the day before they were to receive gynecological surgery and to another group of patients the day before they were to be discharged from the obstetrical service. The hospital patients were matched on age and intelligence with a group of clerks. Of the 65 comparisons made between groups, 21 significant differences were found, generally indicating greater constriction and more self preoccupation in the hospitalized *Ss*. In another study which utilized hospitalized *Ss*, Meyer, Brown, and Levine (1955) found that H-T-P drawings secured before surgery indicated far more regression than was apparent either clinically or in the postoperative drawings. "The contrast between the pre- and postoperative drawings were often so arresting as to cast doubt upon their being the product of the same individual" (p. 431). Abel (1953), however, found conflicting patterns in human figure drawings in patients operated on for correction of facial disfigurement. Some patients made great changes in drawings as a result of corrective surgery, others made little change, while still a third group made dramatic changes in drawings even though no surgery had been performed.

Bellak (1944) made the stressful conditions part of the TAT test administration by criticizing ("These stories are about the worst I've ever heard" etc.) the *S*'s efforts halfway through the testing session. He found that the stories following these remarks were characterized by an increase in aggressive words. Unfortunately, a control group was not used.

Drugs. The effects of a wide variety of drugs have been studied by means of projective devices, particularly the Rorschach. While there is no record of a Rorschach being given under water, there are reports of Rorschachs given under the influence of mescaline, alcohol, sodium amytal, hyoscine, morphine, and oxygen deprivation. In general, these conditions produced limited effects on the test records. Sodium amytal seemed to make *Ss* more willing to cooperate in taking the Rorschach (Wilkins & Adams, 1947) and subsequent responses seemed less bizarre, permitting more detailed descriptions of personality (Kelly, Levine, Pemberton, & Katz, 1941). When *Ss*, given oxygen through a mask, reached a simulated altitude of 16,000 ft., their Rorschachs showed the same basic pattern as tests given under standard conditions, except for a decrease in movement responses and

an increase in color responses (Hertzman & Seitz, 1942). Movement and color responses showed striking changes when the *S* was under the influence of mescaline (Guttman, 1936). Werthman and Bleuler (1932) have reported that the Rorschach test given under mescaline corresponded more closely to the personality as known in life than the standard Rorschach. They also reported the considerable difficulty in holding the experimental *S* to the task of attending to the Rorschach when he seemed to prefer to attend his own images. The administration of hyoscine to an *S* afflicted with encephalitis produced inconsistent changes in the Rorschach (Cofer, 1947).

While most *Ss* in these experiments were badgered, frustrated and otherwise treated unpleasantly, in three studies *Ss* were served liquor in return for taking a projective test. In one investigation, between 9–15 ounces of 100 proof bourbon were served between administrations of the Rorschach. The conclusions of the experiment support the unsystematic observations of drinkers who have to buy their own: under the influence of alcohol, accuracy of perception, attention to details, self-critical attitudes and self-controlling attitudes decrease (Rabin, Papania, & McMichael, 1954). Studies using the TAT (Tompkins, 1942) and the Rorschach (Kelley & Berrera, 1941) agree that projective responses given when the *S* is intoxicated may vary in minor details but do not differ in any consistent manner from responses given when sober.

Whether Brown's (1943) *Ss* were subjected to noxious or pleasant test conditions is a difficult distinction to make, since he administered morphine to 22 postaddict patients who had previously been abstinent from the drug for a minimum of six months. If the *Ss* did not experience the "right feeling" from the first experimental injection, the dose was increased until a euphoric state was reached. It was reported that intellectual control, originality, and organizational energy were not affected by morphine, but there was a shifting to greater imaginative living. There was no report on subsequent attitudes by the *Ss* toward the drug or this particular method of administering the Rorschach.

Hypnosis. The use of hypnosis with resistant test *Ss* was reported by Wilkins and Adams (1947) who found that this method produced cooperation and a useful test record. One *S*, studied intensively by having seven different moods induced while under hypnosis, was variously instructed that she was grief stricken, euphoric, concerned about an undiagnosed illness, depressed, carrying on an extramarital affair, etc. While the content of the Rorschach tests given immediately after the instructions changed from role to role, the similarity between records was great enough to show that the same underlying personality was being observed (Levine, Grassi, & Gerson: 1943, 1944). A similar procedure was employed by Sarbin (1939) who told his *S* in one hypnotic session that she was Madame Curie and in another session that she was Mae West. A third Rorschach was given while the *S* was under hypnosis without any specific suggestion. It was found that the *S* tended to adopt the set imposed by the hypnotic instructions, so that half of her responses in the Madame Curie instructions concerned science while two thirds of her responses in the Mae West set consisted of clothing and costume associations. The control Rorschach, given in the waking state, contained only two responses found in the hyp-

notic trance states. Hypnosis has also been used to study age regression (Bergmann, Graham, & Leavitt, 1947). The test responses of a 20-year-old *S*, given the Rorschach for the eight alternate years between 3 and 20, seemed to reflect the various stages in the development of his personality.

Two studies used the Rorschach with *S*s who had been instructed under hypnosis to feel hostile toward the examiner (Counts & Mensh, 1950; Pattie, 1954). To illustrate the difficulties involved in using hypnosis for this purpose, the conclusions in one study were different from those in the other. Counts and Mensh (1950) found that the hostility to the *E* which appeared during the psychiatric interview was not reflected in the Rorschach, while Pattie (1954) could divide his 14 *S*s into 3 groups, one of which ($N=8$) showed a two-fold increase in hostile content in the posthypnotic Rorschach. The number of white space responses did not increase under the hostile instructions. A considerable degree of individual differences was also found by Arluck and Balinsky (1953) who used as the hypnotic suggestion "you are mature, warm and outgoing," etc. For some *S*s the responses on both the Rorschach and a sentence completion test showed numerous differences with prehypnotic test records, but for others the two sessions produced highly similar results.

Qualitative and quantitative differences following the hypnotic induction of an elated mood and a despondent mood were found on a word association test (Fisher & Marrow, 1934). The reactions times were longest for unpleasantly toned words and were fastest for neutral words.

Special training and experience. Several experiments have investigated the influence of perceptual training upon Rorschach scores. Knopf (1954) provided pre-Rorschach perceptual training in finding animal or animal parts for one group of *S*s while the second group watched a film on the nature of color. He concluded that the over-all picture of the personality remained basically unchanged. Kurtz and Riggs (1954) similarly found no differences in group Rorschach scores in *S*s who had first been exposed to a visual set to perceive animals. "So far as this study is concerned, Rorschach workers remain secure in the assumption that implicit peripheral sets will not influence test results to any appreciable extent" (p. 469). Nor did Norman, Leverant, and Redlo (1952) find that Rorschach scores were altered by having one group of *S*s first look at colored food ads while another group looked at pictures of people in motion. Evidence that perceptual training can influence Rorschach performance has been reported by Keyes (1954) and Leventhal (1956). Subjects trained on stimuli similar to the Street Gestalt pictures produced an increase in the number of whole responses on the group Rorschach (Keyes, 1954). Training on the Gottschaldt figures before an administration of the group Rorschach resulted in lower *W* and *Z* scores (Leventhal, 1956).

Giving children a "gratifying" experience prior to testing seemed to improve performance on the Draw-A-Person test (Reichenberg-Hackett, 1953). Coleman (1947) showed a neutrally toned film to children the evening before they took the second half of the TAT. Of the 370 stories told after the film only one clearly reflected the content of the film. Kimble (1945) gave the Rorschach test twice, once in the standard office procedure and once in the college cafeterias with at least two other people

present. The test results were quite similar in the two situations, the most important difference occurring in the experience balance, more color responses being given in the social situation. Gibby, Stotsky, and Miller (1954) concluded that the experience of taking a psychological test (Bender-Gestalt, TAT, Wechsler-Bellevue, or Goldstein-Scheerer) immediately before an administration of the Rorschach did not alter any of 11 selected Rorschach variables.

Deprivation or arousal of needs. The effects of hunger on associations to ambiguous stimuli have received the attention of three investigators. School children showed more food associations on a word association test shortly before a regular meal than on the test given after the meal (Sanford, 1936). Undergraduates who had fasted for 24 hours before taking five projective tests gave more food responses than did *Ss* who had abstained from food for shorter periods of time. The increase was not a straight line function of time, however, the most hungry *Ss* giving only slightly more food responses than the groups examined near the end of a normal eating cycle (Sanford, 1937). Atkinson and McClelland (1948) were able to derive a food score on TAT stories that could differentiate reliably among groups deprived of food for 1, 4, or 16 hours. While there was no increase in food imagery in these stories, there was a marked increase in food deprivation themes. While the TAT may be able to reflect a state of food deprivation the report of Franklin and Brozek (1949) indicated that the Rosenzweig P-F test responses of a group of conscientious objectors did not change significantly from a period of semi-starvation to a state of nutritional rehabilitation.

Two studies investigated the effects of motor inhibition on the production of movement responses in the Rorschach test. Korchin, Meltzoff, and Singer (1951) restricted normal motor behavior in their *Ss* by having them write a standard phrase as slowly as possible. Singer, Meltzoff, and Goldman (1952) reduced the motor behavior of *Ss* by requesting them to stand for five minutes without moving. In each experiment, the *Ss* denied motor activity gave more movement responses than a control group. However, Singer et al. (1952) also found that increasing motor behavior (by having *Ss* engage in five minutes of vigorous calisthenics) did not reduce the number of movement responses.

The relationship of experimentally aroused needs on projective test responses has also been investigated. When the need for achievement was induced by means of ego-involving instructions, stories of college men increased in achievement imagery themes and themes of instrumental acts and attitudes related to achievement (McClelland, Clark, Roby, & Atkinson, 1949). The stories told by high school boys to pictures of male figures also showed an increased number of themes related to achievement as a result of special achievement arousing conditions. These achievement arousing conditions, however, were not effective in increasing achievement themes of either high school girls or college girls (Veroff, Wilcox, & Atkinson, 1953).

A sociometric test given prior to the administration of the TAT seemed to increase themes related to the affiliation motive (Shipley & Veroff, 1952; Atkinson, Heyns, & Veroff, 1953). Freshmen who were rejected for membership in a fraternity gave more affiliation themes on the TAT than students who were accepted into fraternities (Shipley & Veroff, 1952).

Clark (1952) attempted to raise the level of sexual motivation by showing *Ss* slides of nude females before administering the TAT. A control group was shown slides of landscapes. Under these conditions the group shown the slides of the nudes gave fewer manifest sexual responses and fewer themes of guilt than the control *Ss*. However, when the experiment was repeated after all the *Ss* had participated in a beer party, the experimental group gave more stories of manifest sex and guilt than those *Ss* who had been shown the landscape slides. Rabin, Nelson, and Clark (1954) had one group of undergraduate males waiting to take the Rorschach test remain in a room decorated with anatomical charts and surgical pictures; a second group, somewhat more fortunate in their assignment to experimental conditions, was seated in a room decorated with photographs of nude and seminude females; the control group waited in an undecorated room. While there was no difference between groups in the number of anatomy responses, there was a significant difference in the number of sexual responses.

The Influence of the Examiner

While the evidence regarding situational factors in projective testing has been compiled over the years, the studies dealing with interpersonal influences is of rather recent origin. Guilford's 1947 study was the first in this area and it was three years before the next experiments were reported. The relative neglect of this problem was not entirely due to lack of awareness of its importance, since McFarlane (1942) wrote that "interpretation in the hands of the clinically inexperienced, the doctrinaire, or the methodologically uninformed easily degenerates into nothing more but one more predictive tool—to wit, one which discloses the organizing dynamics of the interpreter rather than the organizing dynamics of the research subject" (p. 405), and Joel warned (1949) that "even if it were possible for the examiner always actually to feel the way he pretends he does, we should not forget that the subject reacts not only to the examiner's real attitude, but also to what he thinks the examiner's attitude is" (p. 480). Probably the greatest deterrents to exploration of this question were the facts that the notion of *E* influence struck at the heart of the X-ray concept of projective tests and the extreme complexity and subtlety of the interpersonal testing situation, so aptly discussed by Schafer (1954), made experimentation difficult.

The first studies tested the hypothesis that *Es* would differ in the responses they elicited from *Ss* by analyzing test records secured from the files of a clinic. As interest turned to determining which characteristics of *Es* were related to differences in *Ss*' responses, such physical attributes of *Es* as skin color, size, and sex were investigated, as well as personality variables revealed in psychological tests, generally the Rorschach. The interaction of *E* and *S* has been studied on several occasions, either by controlling the warm-cold dimension or by contrasting tests taken with *E* present with those taken with *E* absent.

A completely different approach to this problem is through the use of hypnosis. While no study used hypnosis primarily to investigate the testing relationship, most of the experiments using hypnosis do report that *S*'s test behavior varied with the hypnotic suggestion. Thus far only two studies (Gross, 1959; Wickes, 1956) attempted to establish oper-

ant conditioning of *S*'s verbal behavior, but this method seems so promising that undoubtedly it will become more widely used in investigating examiner influence.

The examiner's physical characteristics. The most immediately apparent characteristics of the *E* are his skin color, sex, and body build. Each of these attributes has been investigated for possible influence on the *S*'s responses. Three studies related the sex of *E* to sexual responses on the Rorschach test. Alden and Benton (1951) selected 100 Rorschach records from the files of a VA hospital; all test *S*s were males, 50 of them tested by a female *E* and 50 by a male. There were no significant differences in either overt or covert sexual responses that could be attributed to the sex of *E*. Exactly contradictory results were reported by Curtis and Wolf (1951). Again using the Rorschach record of male veterans, comparisons were made of the overt and covert sexual responses given by 386 *S*s to three female and seven male *E*s. Statistically significant differences were obtained. Rabin et al. (1954) found that sometimes the sex of *E* makes a difference and sometimes it does not. The *S*s who had waited for the Rorschach examination in a room decorated with anatomical charts did not differ in the number of anatomical responses given to the male and female *E*s, but those male *S*s who had waited in a room decorated with pictures of nude women gave significantly more sexual responses to the male *E* than to the female. Clark (1952) found that male *S*s gave more manifest sexual responses and more guilt responses on the TAT to a male *E* than to an attractive, rather seductive, female *E*.

The influence of the *E*'s size and sex on Draw-A-Person productions were investigated by Holtzman (1952) who found that none of his 12 judges could guess better than chance either the sex or the identity of the *E*s by inspecting the drawings of 40 male and 40 female *S*s. Two male *E*s, one of whom was nearly a foot taller and 60 pounds heavier than the other, and two female *E*s differing in "degree of feminine qualities," were used. Garfield, Blek, and Melker (1952) used two female and two male *E*s to administer the TAT to 54 male and 56 female *S*s. Neither the sex of the *E* nor the interaction of the sex of *E* and *S* produced significant differences in the stories.

Riess, Schwartz, and Cottingham (1950) investigated the responses of negro and white *S*s to Negro and white stimulus figures on TAT cards administered by a Negro and white *E*, and concluded that skin color of the *E* did not affect the length of stories. While most investigators look for the influence of the *E* in verbalized test responses, Rankin and Campbell (1955) used as their dependent variable the galvanic skin response of male *S*s to a Negro and white *E*. In a word association situation the *E* checked and adjusted on four occasions dummy equipment connected to *S*'s left wrist. It was found that there was a higher differential galvanic skin response to the Negro *E*, but since he was 9 years older, 2½ inches taller and 27 pounds heavier than the white *E*, there was no conclusive proof that the difference was a function of skin color.

Presence or absence of examiner from the testing room. Bernstein (1956) found that TAT stories written in the absence of the *E* more frequently contained sad themes, sad outcomes, and showed greater *S* involvement than stories written with the *E* present in the testing room. Certain aspects of H-T-P drawings secured from applicants for employ-

ment at a state school were found to be a function of the *E*'s presence when the test was taken. Total size of the drawing, house size, house features, and person features all differed significantly between the group taking the test with *E* present and the group taking the test with the *E* absent (Cassel, Johnson, & Burns, 1958). Van Krevelen (1954b) tested 20 *S*s with 2 cards of the MAPS series; one story was dictated to her and the other was written in her absence. The only significant difference between the two groups was that the more ambiguous of the two cards produced more words in the written, *E*-absent situation. The absence of the *E* seemed to have more effect on Szondi test results (Van Krevelen, 1954a). The *S*s were sometimes administered the Szondi by the *E* and at other times took the test themselves. When *E* was absent, *S*s showed greater consistency, demonstrated more plus-minus reactions, and had a greater sum of open and plus-minus reactions than when *E* was present.

Warm-Cold examiner behavior. Luft (1953) varied the interaction between *E* and *S* by acting warm and friendly to some *S*s and cold and blunt to others. The cold interaction consisted of asking the *S*s their social security number and draft status and by giving a short quiz on current events (e.g., "Which horse won the Kentucky Derby?") before administering 10 homemade ink blots. When the *S*s were asked which inkblots they liked and which they disliked, the group treated in the warm fashion indicated that they liked a mean of 7.6 blots, while the cold *S*s liked only a mean of 3.1 blots, a difference significant beyond the .001 level. Lord (1950) used three styles of administering the Rorschach—neutral, positive, and negative—and three female *E*s. In the positive interaction *E* was instructed to look at *S* with a smile and to be warm and charming; the negative interaction called for *E* to assume the role of a harsh, demanding, authoritative figure, deliberately unconcerned about *S*. Each *S* took the Rorschach three times with the order of *E* and administration counterbalanced. As a result of the different methods of interaction, the protocols elicited from the warm administration produced more responses, more evidence of intellectual and creative imagination, less indication of sterotyped thinking and increased evidence of greater ease in interpersonal relations. In the cold administration, responses indicating imaginative, creative thinking were reduced, there appeared to be a withdrawal from emotional stimuli and there was a rise in self-questioning feelings.

Operant conditioning of subject's verbal behavior. Two studies investigated the extent to which *S*'s associations to ink blots could be determined by the *E*'s behavior. Wickes (1956) used 30 homemade inkblots, two *E*s, and 36 undergraduate *S*s divided into two experimental groups and one control group. In one experimental group the first 15 cards were given in the standard manner; with Card 16 the *E* said, "fine," to the first movement response, "good," to the second movement response, and "all right" to the third movement response, making these comments in regular sequence to the end of the testing. In the second experimental group the first 15 cards were given in standard fashion, but with Card 16 the *E* made various postural and gestural changes, nodding his head three times to the first movement response, smiling on the next movement response and leaning forward in the chair after the third

movement response, repeating this sequence to the end of the testing. When the responses on the first 15 and second 15 cards were compared for all groups, it was found that the experimental group given verbal reinforcement made a significant increase (.025 level) in movement responses on the second half of the test, the experimental group given postural reinforcement made a significant increase (.005 level) in movement responses in the second half of the test, while there was no difference in the number of movement responses in the control group. Gross (1959) verbally reinforced ("good") one group of *Ss* for each human response on the Rorschach while for a second group he nodded his head once following each human response. Both the verbally reinforced group and the nonverbally reinforced group produced more human responses than a control group.

The effect of the examiner as a person: no assessment of examiner personality. In a report not seen by the writer, but quoted by Lord (1950) and others, Guilford (1947) concluded that some *Es* elicited more responses from their *Ss* in administering the Rorschach test than other *Es*. Baughman (1951) also noted that certain *Es* seemed to produce more responses in selected Rorschach categories than other *Es*. He selected 633 protocols secured by 15 *Es* from the files of a veteran's outpatient clinic, and found 12 of 22 scoring categories differing significantly at the .001 level, with four additional differences significant at the .05 level. Unfortunately, the protocols were not scored by the investigator, so that the differences found may have resulted from the psychologists' procedures in scoring, rather than from their influence on *S*. Both Wickes (1956) and Bernstein (1956) used two *Es* in their studies and each concluded that the psychologists did not exert any significant influence on the *Ss'* responses.

To control for some of the sources of variation in responding to the Rorschach, Gibby (1952) had 9 *Es* use a standardized inquiry in testing 135 *Ss*. Despite this, significant differences in responses were found for 6 of the 11 determinants investigated. "The stimulus value of the examiner therefore must be considered a factor which influences the inquiry responses of the subject and therefore the final Rorschach psychogram. Standardization of the inquiry does not eliminate examiner differences" (p. 452). An attempt was made by Gibby, Miller, and Walker (1953) to secure a homogeneous group of patients whose Rorschach records could be analyzed for *E* influence. All *Ss* used were male veterans, white, 25–32 years old, had functional rather than organic ailments, and were the most recent patients tested by *E*. All 12 *Es* whose records were analyzed had a minimum of two years experience, used Beck's Rorschach method and had tested at least 20 *Ss* who met the criteria for the study. All protocols were coded and scored blindly. Of the nine absolute scores which were investigated, three were significant at the .05 level or better, and of the 8 percentage scores, 3 were significant at the .05 level or better. The investigators concluded that "there are significant over-all differences in the determinants obtained by various examiners from comparable groups of subjects. . . . It is also probable that certain examiners tend to obtain successively dysphoric records while others with comparable patients will seldom elicit such reactions" (p. 426).

Robinson and Cohen (1954) examined the case reports prepared by three psychological interns. The last

30 reports for each of the three interns were examined for the variables of dependence, independence, aggression, and abasement. When the incidence of these variables in the case reports was compared for each of the three psychologists, 6 of the 12 comparisons were significant at the .05 level. When components of these variables were considered, 12 of the 24 comparisons were significant at the .05 level of confidence. The investigators concluded that this study raises "a serious question about the objectivity of methods of evaluation and prediction if they must rely solely upon psychological reports for their basis" (p. 335).

The effect of the examiner as a person; examiner's personality assessed. Hammer and Piotrowski (1953) asked three staff psychologists and three interns to rate 400 H-T-P drawings on a 3-point scale of aggression. The clinicians were themselves rated by one of the investigators on the degree of aggression and hostility they manifested in dealing with patients and staff members. In addition, the clinicians also took the Szondi test which was scored by Susan Deri for degree of hostile and aggressive impulses. A rank order correlation of .94 was found for the degree of hostility the clinicians saw in the H-T-P productions and the evaluation made of their interpersonal hostility. The rating of the clinicians' Szondi tests also yielded a rank order correlation of .94 with their evaluations of the hostility in the H-T-P. The authors concluded that "just as a subject's performance on a projective technique is a function of his personality, his needs, conflicts, desires and past experiences, so too, although to a lesser degree, is the interpretation of a projective protocol influenced by the personality pattern of the interpreter" (p. 214–215). Filer (1952) analyzed 156 case reports submitted by 13 male clinicians for references to hostility, hostility turned inward, passive dependency, and feelings of inferiority. Evaluations were also obtained of the clinician's behavior in terms of ascendency, depression, intropunitiveness, extrapunitiveness, and impunitiveness. The results indicated a complex relationship between the judgments of the clinician and the judgments made by the clinician. For example, *E*s who stressed hostility turned inward were rated depressed and intrapunitive. A separate study of references to defense mechanisms indicated that the three most frequently mentioned mechanisms were more characteristic of *E* than the test S.

Berger (1954) used the personal Rorschach results of eight VA trainees to compare with the Rorschach records these *E*s elicited from VA patients. Contrary to the findings of most other studies, Berger found no significant differences in *E* influence on 12 variables in the patients' Rorschachs. However, when both the *E* and his *S*s were rank-ordered on the 12 variables, a rho correlation of .86 was found for the number of popular responses, a rho of .80 for white space responses and a rho of $-.54$ for *Y* responses.

Lord (1950) did not attempt to make a formal assessment of the personalities of her three female *E*s, but she obtained descriptions of them at a subjective, intuitive level from two clinicians, relating these descriptions to the Rorschach responses secured from *S*s. The first *E*, described by the judges as a cold, inflexible, masculine, castrating woman, produced Rorschachs from her *S*s that made it appear that they had faced a threatening, frustrating situation. The second *E*, described as the most feminine of the group, the softest and

most mother-like, elicited Rorschach records that suggested her *S*s had been under controlled excitement without great anxiety or tension and had not been greatly stimulated intellectually. The third *E* was the oldest of the three and was described as the most flexible, exuberant, sympathetic, intermediate in feminine qualities. Her *S*s' Rorschach records suggested that they had been challenged and made anxious, but there was also evidence of easy rapport and a relative absence of controlling devices.

The TAT and MMPI records of nine *E*s were evaluated and compared with the Rorschach responses the *E*s had elicited. An analysis of *E* variance showed 22 of 37 selected Rorschach variables significant at the .05 level or better (Miller, Sanders, & Cleveland, 1950). Sanders and Cleveland (1953) trained nine second year graduate students to administer the Rorschach test after first obtaining a personal Rorschach from each. After a period of training, each *E* administered 20 Rorschachs to undergraduate *S*s; at the end of each testing session *S* was asked to rate *E* on measures of overt anxiety and hostility. Indications of *E*s' covert anxiety and hostility were obtained by rating their personal Rorschach records. When the number of responses given by each *S* was held constant it was found that of the 20 variables investigated, 9 differed significantly among *E*s. The *E*s rated high on overt anxiety by their *S*s elicited more responses, more white space responses and more color responses than those rated low in overt anxiety. No differences in *S*s' scores could be accounted for by scores of the *E*s on covert anxiety. The *E*s rated high on hostility by their *S*s elicited greater *Y*% and *A*% in the Rorschachs they administered and secured less hostile and human content than those *E*s whose *S*s rated them low in hostility. The *S*s' reaction times seemed slower when *E* was perceived as hostile, a finding in agreement with the observation made of the test behavior of *S*s with hypnotically induced hostility to *E* (Counts & Mensh, 1950; Pattie, 1954). When *E*s' personal Rorschachs were examined for hostility, it was found that the more hostile *E*s elicited less *Y*%, less *A*%, more hostile content and more human content than those *E*s with low covert hostility. The authors concluded that "The results of this study provide evidence that a subject's responses on the Rorschach test are not solely a product of his own emotional problems and personality structure" (p. 48). While the investigators clearly indicated that *E* as a person can influence the test results, they estimated that the extent of this influence can account for only 3–7% of the total variance in the Rorschach scores.

The Influence of the Subject

While there is wide theoretical agreement that each party in the testing situation exerts an influence on the other, the experimental evidence is limited almost entirely to the effect *E* has on *S*. Perhaps the chief reason for the failure to investigate the manner in which *S* can influence the psychologist is the lack of traditional experimental procedures that can afford control of the *S*'s behavior while the *E*'s behavior is allowed to vary. The one study in this area (Masling, 1957) controlled *S*'s behavior by using attractive female accomplices who posed as test *S*s, acting warm or cold to *E*. The dependent variable was the interpretation placed on sentence completion protocols by eight graduate student *E*s. It was found that when *S* acted

warm to *E* her protocol was interpreted more favorably (i.e., she was seen in better mental health) than when she acted cold. In addition, the results indicated that when *E* saw two *Ss*, one of whom was cold and the other warm, the protocol of the warm *S* was interpreted more favorably than that of the cold *S*.

Discussion

The studies presented in this paper have been reviewed rather uncritically, with the emphasis on content rather than on adequacy of experimental design. Since faulty experimental procedures appear with regularity in the studies in this area, it might be worthwhile to examine more closely the more commonly found limitations in design:

1. No study reviewed here extensively sampled the *E* population. As Hammond (1954) has indicated, representative design demands that both *E* and *S* populations be adequately sampled if generalizations are to be made to larger groups of *S* and *E*. Most studies cited here, however, utilized only one *E*, with only Baughman (1951) using as many as 15 *Es*. The general results of the work on *E* differences makes clear how tenuous it is to assume that one *E* is drawn from the same population as any other *E*.

2. Not only has the *E* population been inadequately sampled, but the little attention given to *Es* has been directed for the most part to graduate students. While it is legitimate to work with a graduate student population it is inappropriate to generalize findings to a population of older, more experienced *Es*.

3. Those studies which investigated the influence of *E* differences by utilizing a random sample of cases found in the files of a clinic, make the assumption as Levy (1956) has indicated, that the cases were originally assigned on a random basis. This assumption may not always be valid, due to differences in *E* schedules, interests, and competence. As a result, differences in test records may be in part a function of uncontrolled bias in the selection of *Ss*. It is far better procedure for the investigator to control the assignment of cases than to assume existing cases had been randomly assigned.

4. A frequent method of assessing *E* personality has been to ask the *E* to take a psychological test. As the results of this review make quite clear, the orientation *S* has toward the test considerably influences his responses. Few graduate students in psychology are naive regarding the more common projective tests, even if they have never seen them before. The meaning of a Rorschach test taken by a graduate student, therefore, is unclear and cannot be easily related to differences in *Ss*' responses. A better way of evaluating *E*'s personality might be to obtain judgments by his supervisors and colleagues. Another method that shows promise is that used by Sanders and Cleveland (1953), who asked the test *Ss* to make ratings of their impressions of *E*.

5. Most investigations of susceptibility of *Ss*' responses to situational influence have been conducted empirically, with no prior attempt made to predict where differences would be found. Research on the Rorschach has been particularly culpable in this regard. Since the Rorschach is still primarily an empirically, rather than theoretically, based instrument, most investigators have attempted to determine only if differences would occur between experimental and control groups, but, on finding differences, have been unable to interpret their meaning. As a result of this ap-

proach, almost every Rorschach score has at one time or another been found to be a function of some experimental variable: *Z*, *W*, *W%*, *Ds*, *D*, *Dd%*, *F−*, *F+%*, *F%* *FM*, *m*, *M*, *CF*, *C*, *Y*, *Y%*, *A*, *A%*, *P*, *R*, reaction time and experience balance have all been reported to change as a result of experimental conditions.

6. Many of the Rorschach studies appearing before Cronbach's (1949) critique of Rorschach research did not control for the number of responses, but assumed that all differences in determinants could be attributed to the experimental variable. To a large extent investigators now attempt to partial out differences in the number of responses, but an occasional study will still disregard this factor. While most of the other statistical errors Cronbach discussed occur far less frequently in later research, inflation of probability levels continues to be a major source of error.

Despite these flaws in design the studies cited here presented strong evidence of situational and interpersonal influences in projective testing. It is important to note, however, that the projective response did not change with any and all conditions imposed by the *E*. The use of drugs produced only minimal changes in the protocol and there was conflicting evidence regarding the importance of such physical differences among *E*s as skin color, sex and size. What appeared to be the crucial element was the extent to which *S*'s attitude toward the *total* testing situation was influenced by the experimental conditions. Where the experimental variable was peripheral to the examination, as in Coleman's (1947) showing of a film the night before administering the TAT, no appreciable effect was introduced in the protocol; when the condition was sufficiently unique and sufficiently contiguous to the testing session, as in the use of a waiting room decorated with pictures of nude women (Rabin, Nelson, & Clark, 1954), *S*s evidently construed the experimental conditions to be a part of the total testing situation.

There is considerable evidence that *S*s in an unstructured situation will utilize all available cues to complete their assigned task. The *S* in the projective test setting will not only use those cues furnished by the ink blot or picture, but also those supplied by his feelings about the examiner, those furnished by his needs, attitudes and fears, those implied in the instructions, the room, and previous knowledge of the test, and those cues supplied consciously or unconsciously by *E*. When *E* faces the ambiguous situation of supplying meaning to a series of isolated, discrete responses, he will not only rely on *S*'s responses, but also on those cues furnished by his training and theoretical orientation, his own needs and expectations, his feelings about *S* and the constructions he places on *S*'s test behavior and attitudes. In short, these studies demonstrate that *E* and *S* behave as we should expect, considering our knowledge of behavior in ambiguous settings.

Thus, the procedure that many clinicians hoped would serve as an X ray proves, on close examination, to function also as a mirror, reflecting impartially *S*, *E*, the situation and their interactions. This need not be a cause for despair, except for those who feel that *E* and situational influences contaminate a protocol. These influences are not sources of error, however, but indications of adaptation to the task. One reason for the poor record of blind analysis as a procedure for validating projective devices is that this method can utilize only a fraction of the material avail-

able in a protocol. Instead of trying to eliminate interpersonal and situational influence *E* might better make a more thorough search of his own attitudes and of *S*'s attitudes toward the test and the situation (Leventhal, Rosenblatt, Gluck, & Slepian, 1958). The interpersonal situation "is not an evil. It should not be striven against. As in psychoanalytic technique, this relationship must be regarded as inevitable, as a potentially significant influence on the patient's productions, and as a possible goldmine of material for interpretation" (Schafer, 1954, p. 6).

The important problems in this area remain unsolved. What effect does experience have on *E*'s sensitivity to the attitudes of *S*? Of what importance is the psychological health of *S* in his response to the attitudes of *E*? Little is understood about the circumstances which prompt an *S* to rely heavily on interpersonal cues, nor do we know much of the forces acting on an *E* who is faced with a belligerent, or overly cooperative or suspicious *S*. Most important of all, the great bulk of the studies cited here indicate only that situational and interpersonal variables influence the projective protocol, without in any way describing how these variables impinge on *E* and *S*. How *S* senses that *E* is hostile and how *E* realizes *S* is trying to control him can only be determined by studying the interaction process itself. It is interesting to note that no study reviewed here used verbatim recordings of the testing session, despite the growing popularity of verbatim transcripts in research on psychotherapy. Hopefully, future research on projective testing will investigate more fully the step-by-step transactions between *E* and *S*.

REFERENCES

ABEL, T. M. Figure drawings and facial disfigurement. *Amer. J. Orthopsychiat.*, 1943, **23**, 253–264.

ABRAMSON, L. S. The influence of set for area on the Rorschach test results. *J. consult. Psychol.*, 1951, **15**, 337–342.

ALDEN, PRISCILLA, & BENTON, A. L. Relationship of sex of examiner to incidence of Rorschach responses with sexual content. *J. proj. Tech.*, 1951, **15**, 230–234.

ARLUCK, E. W., & BALINSKY, B. Possible shifts in functioning through hypnotic suggestion. *J. proj. Tech.*, 1953, **17**, 447–454.

ATKINSON, J. W., HEYNS, R. W., & VEROFF, J. Effect of experimental arousal of the affiliation motive on thematic apperception. *Amer. Psychologist*, 1953, **8**, 313–314. (Abstract)

ATKINSON, J. W., & MCCLELLAND, D. C. The projective expression of needs: II. The effect of different intensities of the hunger drive on thematic apperception. *J. exp. Psychol.*, 1948, **38**, 643–658.

BAUGHMAN, E. E. Rorschach scores as a function of examiner differences. *J. proj. Tech.*, 1951, **15**, 243–249.

BELLAK, L. The concept of projection: An experimental investigation and study of the concept. *Psychiatry*, 1944, **7**, 353–370.

BERGER, D. Examiner influence on the Rorschach. *J. clin. Psychol.*, 1954, **10**, 245–248.

BERGMANN, M. S., GRAHAM, H., & LEAVITT, H. C. Rorschach exploration of consecutive hypnotic chronological age level regression. *Psychosom. Med.*, 1947, **9**, 20–28.

BERNSTEIN, L. The examiner as inhibiting factor in clinical testing. *J. consult. Psychol.*, 1956, **20**, 287–290.

BROWN, R. R. The effect of morphine upon the Rorschach pattern in post-addicts. *Amer. J. Orthopsychiat.*, 1943, **13**, 339–343.

CALDEN, G., & COHEN, L. B. The relationship between ego-involvement and test definition to Rorschach test performance. *J. proj. Tech.*, 1953, **17**, 300–311.

CARP, A. L., & SHAVZIN, A. R. The susceptibility to falsification of the Rorschach diagnostic technique. *J. consult. Psychol.*, 1950, **3**, 230–233.

CASSEL, R. H., JOHNSON, ANNA, & BURNS, W. H. Examiner, ego-defense and the H-T-P test. *J. clin. Psychol.*, 1958, **14**, 157–160.

CLARK, R. A. The projective measurement of experimentally induced levels of sexual mo-

tivation. *J. exp. Psychol.*, 1952, **44**, 391–399.

COFER, C. N. Psychological test performance under hyoscine: A case of post-infectious encephalopathy. *J. gen. Psychol.*, 1947, **36**, 221–228.

COFFIN, T. E. Some conditions of suggestion and suggestibility: A study of certain attitudinal and situational factors influencing the process of suggestion. *Psychol. Monogr.*, 1941, **53**(4, Whole No. 241).

COLEMAN, W. The Thematic Apperception Test I. Effect of recent experiences II. Some quantitative observations. *J. clin. Psychol.*, 1947, **3**, 257–264.

COUNTS, R. M., & MENSH, I. N. Personality characteristics in hypnotically induced hostility. *J. clin. Psychol.*, 1950, **6**, 325–330.

CRANDALL, V. S. Induced frustration and punishment reward expectancy in thematic apperception stories. *J. consult. Psychol.*, 1951, **15**, 400–404.

CRONBACH, L. J. Statistical methods applied to Rorschach scores. *Psychol. Bull.*, 1949, **46**, 393–429.

CRONBACH, L. J. Assessment of individual differences. In P. R. Farnsworth (Ed.) *Annual review of psychology.* Stanford: Ann. Rev., 1956.

CURTIS, H. S., & WOLF, ELIZABETH. The influence of the sex of the examiner on the prediction of sex responses on the Rorschach. *Amer. Psychologist*, 1951, **6**, 345–346. (Abstract)

EICHLER, R. M. Experimental stress and alleged Rorschach indices of anxiety. *J. abnorm. soc. Psychol.*, 1951, **46**, 344–355.

FABRIKANT, B. Rigidity and flexibility on the Rorschach. *J. clin. Psychol.*, 1954, **10**, 255–258.

FILER, R. M. The clinician's personality and his case reports. *Amer. Psychologist*, 1952, **7**, 336. (Abstract)

FISHER, V., & MORROW, A. Experimental study of moods. *Charact. & Pers.*, 1934, **2**, 201–208.

FOSBERG, I. A. Rorschach reactions under varied instructions. *Rorschach Res. Exch.*, 1938, **3**, 12–30.

FOSBERG, I. A. An experimental study of the reliability of the Rorschach psycho-diagnostic technique. *Rorschach Res. Exch.*, 941, **5**, 72–84.

FOSBERG, I. A. How do subjects fake results on the Rorschach test? *Rorschach Res. Exch.*, 1943, **7**, 119–121.

FRANK, L. K. Projective methods for the study of personality. *J. Psychol.*, 1939, **8**, 389–413.

FRANKLIN, J. C., & BROZEK, J. The Rosenzweig P-F test as a measure of frustration response in semistarvation. *J. consult. Psychol.*, 1949, **13**, 293–301.

FRENCH, R. L. Changes in performance on the Rosenzweig Picture-Frustration Study following experimentally induced frustration. *J. consult. Psychol.*, 1950, **14**, 111–115.

GARFIELD, S., BLEK, LIBBY, & MELKER, F. The influence of method of administration and sex differences on selected aspects of TAT stories. *J. consult. Psychol.*, 1952, **16**, 140–145.

GIBBY, R. G. The stability of certain Rorschach variables under conditions of experimentally induced sets: I. The intellectual variable. *J. proj. Tech.*, 1951, **15**, 3–26.

GIBBY, R. G. Examiner influence on the Rorschach inquiry. *J. consult. Psychol.*, 1952, **16**, 449–455.

GIBBY, R. G., MILLER, D. R., & WALKER, E. L. The examiner's influence on the Rorschach protocol. *J. consult. Psychol.*, 1953, **17**, 425–428.

GIBBY, R. G., STOTSKY, B. A., & MILLER, D. R. Influence of the preceding test on the Rorschach protocol. *J. consult. Psychol.*, 1954, **18**, 463–464.

GROSS, L. Effects of verbal and nonverbal reinforcement in the Rorschach. *J. consult. Psychol.*, 1959, **23**, 66–68.

GUILFORD, J. P., & LACEY, J. I. (Eds.) Printed classification tests. (*AAF Aviat. Psychol. Program Res. Rep.* No. 5.) Washington: U. S. Government Printing Office, 1947.

GUTTMAN, E. Artificial psychoses produced by mescaline. *J. ment. Sci.*, 1936, **52**, 203–221.

HAMMER, E. F., & PIOTROWSKI, Z. A. Hostility as a factor in the clinician's personality as it affects his interpretation of projective drawings (H-T-P), *J. proj. Tech.*, 1953, **17**, 210–216.

HAMMOND, K. R. Representative vs. systematic design in clinical psychology. *Psychol. Bull.*, 1954, **51**, 150–159.

HENRY, EDITH, & ROTTER, J. B. Situational influence on Rorschach responses. *J. consult. Psychol.*, 1956, **6**, 457–462.

HERTZMAN, M., & SEITZ, C. Rorschach reactions at high altitudes. *J. Psychol.*, 1942, **14**, 245–257.

HOLTZMAN, W. H. The examiner as a variable in the Draw-A-Person Test. *J. consult. Psychol.*, 1952, **16**, 145–148.

HUTT, M., GIBBY, R. G., MILTON, E. O., & POTTHARST, K. The effect of varied experimental "sets" upon Rorschach test performance. *J. proj. Tech.*, 1950, **14**, 181–187.

JOEL, W. The interpersonal equation in pro-

jective methods. *Rorschach Res. Exch.*, 1949, **13**, 479–482.

KELLEY, D., & BARRERRA, S. Rorschach studies in acute experimental alcoholic intoxication. *Amer. J. Psychiat.*, 1941, **97**, 1341–1364.

KELLEY, D., LEVINE, KATE, PEMBERTON, W., & KATZ, K. Intravenous sodium amytal medication as an aid to the Rorschach method. *Psychiat. Quart.*, 1941, **15**, 68–73.

KEYES, E. J. An experimental investigation of some sources of variance in the whole response to the Rorschach ink blots. *J. clin. Psychol.*, 1954, **10**, 155–160.

KIMBLE, G. A. Social influence on Rorschach records. *J. abnorm. soc. Psychol.*, 1945, **40**, 89–93.

KLATSKIN, ETHELYN. An analysis of the test situation upon the Rorschach record: Formal scoring characteristics. *J. proj. Tech.*, 1952, **16**, 193–199.

KNOPF, I. J. The effects of recent perceptual training and experience on Rorschach performance. *J. clin. Psychol.*, 1954, **10**, 52–56.

KORCHIN, S., MELTZOFF, J., & SINGER, J. L. Motor inhibition and Rorschach movement responses. *Amer. Psychologist*, 1951, **6**, 344–345. (Abstract)

KURTZ, JOSEPHINE, & RIGGS, MARGARET. An attempt to influence the Rorschach test by means of a peripheral set. *J. consult. Psychol.*, 1954, **18**, 465–470.

LEVENTHAL, H. The influence of previous perceptual experience on the variance of the Rorschach *W* and *Z* scores. *J. consult. Psychol.*, 1956, **20**, 93–98.

LEVENTHAL, T., ROSENBLATT, B., GLUCK, M., & SLEPIAN, H. The use of the psychologist-patient relationship in individual diagnostic testing. *Amer. Psychologist*, 1958, **13**, 345. (Abstract)

LEVINE, KATE, GRASSI, J. R., & GERSON, M. J. Hypnotically induced mood changes in the verbal and graphic Rorschach: A case study. *Rorschach Res. Exch.*, 1943, **7**, 130–144.

LEVINE, KATE, GRASSI, J. R., & GERSON, M. J. Hypnotically induced mood changes in the verbal and graphic Rorschach: Part II: The response records. *Rorschach Res Exch.*, 1944, **8**, 104–124.

LEVY, L. H. A note on research methodology used in testing for examiner influence in clinical test performance. *J. consult. Psychol.*, 1956, **20**, 286.

LINDZEY, G. An experimental examination of the scapegoat theory of prejudice. *J. abnorm. soc. Psychol.*, 1950, **45**, 296–309. (a)

LINDZEY, G. An experimental test of the validity of the Rosenzweig Picture-Frustration Study. *J. Pers.*, 1950, **18**, 315–320. (b)

LINDZEY, G., & HERMAN, P. S. Thematic Apperception Test: A note on reliability and situational validity. *J. proj. Tech.*, 1955, **19**, 36–42.

LORD, EDITH. Experimentally induced variation in Rorschach performance. *Psychol. Monogr.*, 1950, **64**(10, Whole No. 316).

LUCHINS, A. S. Situational and attitudinal influences on Rorschach responses. *Amer. J. Psychiat.*, 1947, **103**, 780–784.

LUFT, J. Interaction and projection. *J. proj. Tech.*, 1953, **17**, 489–492.

MCCLELLAND, D. C., CLARK, R. A., ROBY, T. B., & ATKINSON, J. W. The projective expression of needs: IV. The effects of need for achievement on thematic apperception. *J. exp. Psychol.*, 1949, **39**, 242–255.

MCFARLANE, JEAN W. Problems of validation inherent in projective methods. *Amer. J. Orthopsychiat.*, 1942, **12**, 405–410.

MASLING, J. M. The effects of warm and cold interaction on the interpretation of a projective protocol. *J. proj. Tech.*, 1957, **21**, 377–383.

MEYER, B. C., BROWN, F., & LEVINE, A. Observations on the House-Tree-Person drawing test before and after surgery. *Psychosom. Med.*, 1955, **17**, 428–454.

MILLER, D., SANDERS, R., & CLEVELAND, S. E. The relationship between examiner personality and obtained Rorschach protocols: An application of interpersonal relations theory. *Amer. Psychologist*, 1950, **5**, 322–323. (Abstract)

MILLER, D. R. Prediction of behavior by means of the Rorschach test. *J. abnorm. soc. Psychol.*, 1953, **48**, 367–375.

NORMAN, R., LEVERANT, S., & REDLO, MIRIAM. The influence of a superficial immediately preceding "set" upon responses to the Rorschach. *J. consult. Psychol.*, 1952, **16**, 261–264.

PATTIE, F. A. The effect of hypnotically induced hostility on Rorschach performance. *J. clin. Psychol.*, 1954, **10**, 161–164.

RABIN, A., NELSON, W., & CLARK, MARGARET. Rorschach content as a function of perceptual experience and sex of the examiner. *J. clin. Psychol.*, 1954, **10**, 188–190.

RABIN, A., PAPANIA, N., & MCMICHAEL, A. Some effects of alcohol on Rorschach performance. *J. clin. Psychol.*, 1954, **10**, 232–255.

RANKIN, R. E., & CAMPBELL, D. T. Galvanic skin response to negro and white experimenters. *J. abnorm. soc. Psychol.*, 1955, **51**, 30–33.

REICHENBERG-HACKETT, W. Changes in Goodenough drawings after a gratifying experience. *Amer. J. Orthopsychiat.*, 1953, **23**, 501–516.

RIESS, B. F., SCHWARTZ, E. K., & COTTINGHAM, ALICE. An experimental critique of assumptions underlying the negro version of the TAT. *J. abnorm. soc. Psychol.*, 1950, **45**, 700–709.

ROBINSON, J., & COHEN, L. Individual bias in psychological reports. *J. clin. Psychol.*, 1954, **10**, 333–336.

RODNICK, E. H., & KLEBANOFF, S. G. Projective reactions to induced frustration as a measure of social adjustment. *Psychol. Bull.*, 1942, **39**, 489. (Abstract)

SANDERS, R., & CLEVELAND, S. E. The relationship between certain examiner personality variables and subjects' Rorschach scores. *J. proj. Tech.*, 1953, **17**, 34–50.

SANFORD, R. N. The effects of abstinence from food upon imaginal processes: A preliminary experiment. *J. Psychol.*, 1936, **2**, 129–136.

SANFORD, R. N. The effects of abstinence from food upon imaginal processes: A further experiment. *J. Psychol.*, 1937, **3**, 145–149.

SARASON, S. B. *The clinical interaction*. New York: Harper, 1954.

SARBIN, T. R. Rorschach patterns under hypnosis. *Amer. J. Orthopsychiat.*, 1939, **9**, 315–318.

SCHACTEL, E. G. Subjective definitions of the Rorschach test situation and their effect on test performance. Contributions to an understanding of Rorschach's test, III. *Psychiatry*, 1945, **8**, 419–448.

SCHAFER, R. *Psychoanalytic interpretation in Rorschach testing*. New York: Grune & Stratton, 1954.

SCHWARTZ, F., & KATES, S. L. Behn-Rorschach and Rorschach under standard and stress conditions. *J. consult. Psychol.*, 1957, **21**, 335–338.

SHIPLEY, T. E., & VEROFF, J. A projective measure of need for affiliation. *J. exp. Psychol.*, 1952, **43**, 349–356.

SIIPOLA, ELSA, & TAYLOR, VIVIAN. Reactions to ink blots under free and pressure conditions. *J. Pers.*, 1952, **21**, 22–47.

SINGER, J. L., MELTZOFF, J., & GOLDMAN, G. D. Rorschach movement responses following motor inhibition and hyperactivity. *J. consult. Psychol.*, 1952, **16**, 359–364.

SUMMERWELL, HARRIET, CAMPBELL, MARY, & SARASON, I. The effect of differential motivating instructions on the emotional tone and outcome of the TAT stories. *J. consult. Psychol.*, 1958, **22**, 385–388.

TOMPKINS, S. S. The limits of material obtainable in a single case study by daily administration of the TAT. *Psychol. Bull.*, 1942, **39**, 490. (Abstract)

VANKREVELEN, ALICE. Some effects of subject-examiner interaction on projective test performance. *J. proj. Tech.*, 1954, **18**, 107–109. (a)

VANKREVELEN, ALICE. A study of examiner influence on responses to MAPS test materials. *J. clin. Psychol.*, 1954, **10**, 292–293. (b)

VEROFF, J., WILCOX, S., & ATKINSON, J. W. The achievement motive in high school and college age women. *J. abnorm. soc. Psychol.*, 1953, **48**, 108–119.

WALLON, E. J., & WEBB, W. The effect of varying degrees of projection on test scores. *J. consult. Psychol.*, 1957, **21**, 465–472.

WEISSKOFF, EDITH A. The influence of the time factor on Rorschach performance. *Rorschach Res. Exch.*, 1942, **6**, 128–136.

WEISSKOFF, EDITH, & DIEPPA, J. J. Experimentally induced faking of TAT responses. *J. consult. Psychol.*, 1951, **15**, 469–474.

WERTHAM, F., & BLUELER, M. Inconstancy of the formal structure of the personality: Experimental study of the influence of mescaline on the Rorschach test. *Arch. Neurol. Psychiat.*, 1932, **28**, 52–70.

WICKES, T. A. Examiner influence in a test situation. *J. consult. Psychol.*, 1956, **20**, 23–26.

WILKINS, W. L., & ADAMS, A. J. The use of the Rorschach test under hypnosis and under sodium amytal in military psychiatry. *J. gen. Psychol.*, 1947, **36**, 131–138.

(Received February 9, 1959)

PSYCHOLOGICAL BULLETIN
Vol. 57, No. 1, 1960

NOTE ON THE MULTITRAIT-MULTIMETHOD MATRIX

LLOYD G. HUMPHREYS
University of Illinois

The recent article by Campbell and Fiske (1959) is stimulating and sound, but one of their conclusions can be supplemented to indicate a desirable direction for further research. They conclude as follows: "Measures of the same trait should correlate higher with each other than they do with measures of different traits involving separate methods. Ideally, these validity values should also be higher than the correlations among different traits measured by the same method." There is no arguing with their first criterion, and their second is hedged by the use of "ideally." Even so, the second is too strong. It is the purpose of this note to show that the second criterion is better characterized as "nice." The degree of importance to be attached to it is simply a function of the number of different methods that can be used to measure the trait.

The procedure required to handle the problem posed by correlations between measures of different traits, utilizing the same method, that are higher than the correlations between measures of the same trait across different methods is implicit in Spearman. It has also been suggested by several measurement theorists subsequent to Spearman, but has rarely been followed explicitly as a basis for test construction. Yet, considering the present status of personality measurement, one might well conclude that this procedure constitutes our only hope in this area.

Consider the following four rows of a hypothetical rotated factor matrix:

	Trait A	Trait B	Method I	Method II	Method III	Method IV	Method V	Method VI
Test 1	40	00	80	00	00	00	00	00
Test 2	40	00	00	80	00	00	00	00
Test 7	00	40	80	00	00	00	00	00
Test 12	00	40	00	00	00	00	00	00

The correlation between Variables 1 and 2, both assumed to measure the same trait, is only .16, but the correlation between 1 and 7, measuring different traits, is .64. Clearly, the ideal of Campbell and Fiske (1959) is not met, and by a very substantial amount. But now let us assume that Variables 3 through 6 (not shown) also have loadings of .40 on Trait A, and each has a loading of .80 on one of the succeeding methods factors (Factors III through VI). Similarly, Variables 8 through 11 have loadings of .40 on Trait B, with method loadings of .80 on succeeding factors. If we now sum Variables 1 through 6 and Variables 7 through

12, the two aggregates will have the following factor loadings, as determined by the formula for correlations of sums:

	Trait A	Trait B	Method I	Method II	Method III	Method IV	Method V	Method VI
Sum of 1–6	73	00	24	24	24	24	24	24
Sum of 7–12	00	73	24	24	24	24	24	24

The factor loadings for methods of .80 have now shrunk to .24 and the loadings of .40 for traits have increased to .73. The correlation between the two measures of the orthogonal traits is .35, because they have methods variance in common, while the correlation with an equally fallible measure of the same trait is only .53. Even so, we can conclude that our two aggregates have psychometric characteristics which make them potentially useful tools for further research involving these two traits. Both could be further improved by the addition of methods, by the use of nonoverlapping methods, or by the use of suppressors for methods.

Note that the first criterion of Campbell and Fiske must hold if this procedure is to be successful. Beyond that, it depends entirely on our ability to devise enough methods of measurement for each trait in which we are interested. The number of methods required in any particular situation will depend on how low the intercorrelations are when trait is held constant and methods vary, as opposed to how high the intercorrelations are when method is held constant and traits vary.

In discussing so glibly the problem of increasing the number of methods which measure the trait in question, the writer does not intend to convey the impression that the problem is an easy one empirically. Quite the reverse is true, but the difficulty of the search should not deter us. After all, we can hardly point with pride to what we have accomplished in the personality realm in the attempt to find good trait measures defined by single methods.

The reader should not conclude that aggregating across methods to obtain better trait measures is restricted to the personality realm. Systematic application of this procedure to aptitude measurement is also desirable. Guilford's analysis of the ability domain (1956) could serve as the basis for deciding which dimension or dimensions should be added across in constructing socially or scientifically useful trait measures. This is in contrast to a point of view, which is apparently held by Guilford, that recommends a separate test for every cell in his table. Thus, in constructing a reasoning test, it might be desirable to aggregate across methods, e.g., series, analogies, and classification items; and across content, e.g., words, numbers, figures, letters, and objects. The fact that an ability *can* be fractionated is insufficient basis for concluding that it *should* be fractionated.

REFERENCES

CAMPBELL, D. T., & FISKE, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.*, 1959, **56**, 81–105.

GUILFORD, J. P. The structure of intellect. *Psychol. Bull.*, 1956, **53**, 267–293.

(Received April 15, 1959)

APA Members and Journal Subscribers—
Are you going to move?

If you move—

your journals will not follow you from your old address to your new one

When you move—

notify the APA Subscription Office

Formerly, journals that could not be delivered because subscribers had not notified the APA of a new address were reclaimed by the APA, and the postage was remailed to the subscribers at his new address. This was always expensive. Recent changes in the postal laws and regulations have made the expense prohibitive. Undeliverable copies are now destroyed by the Post Office. Subscribers who do not receive a journal because an address change was changed 1¢ regular single issue price for a replacement copy.

So—when you move—

Notify the postmaster at your old address and guarantee that you will pay the forwarding postage.

Notify the APA Member Subscription Department as early as possible—by at least the tenth of the month preceding the month when the change should take effect.

GEORGE BANTA COMPANY, INC., MENASHA, WISCONSIN, U.S.A.